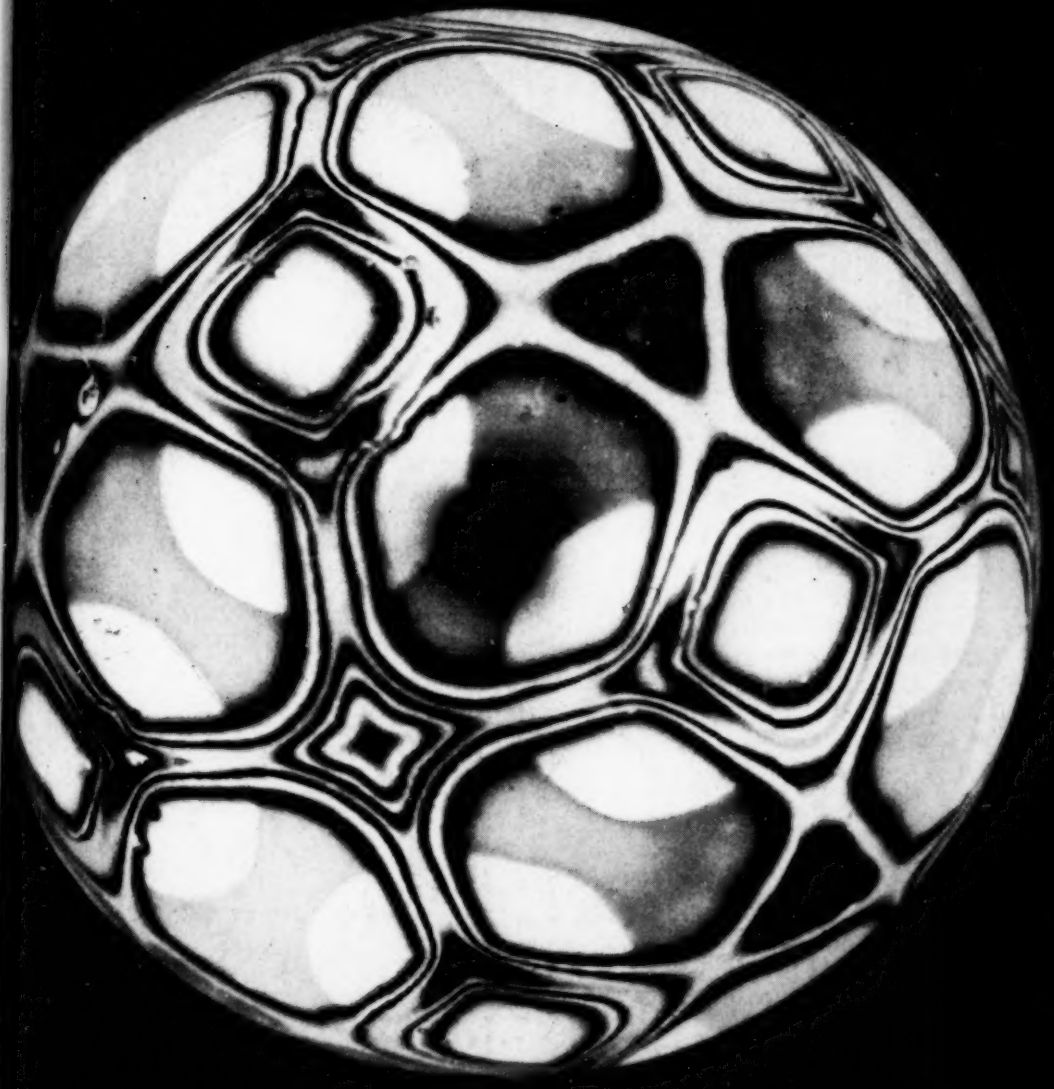
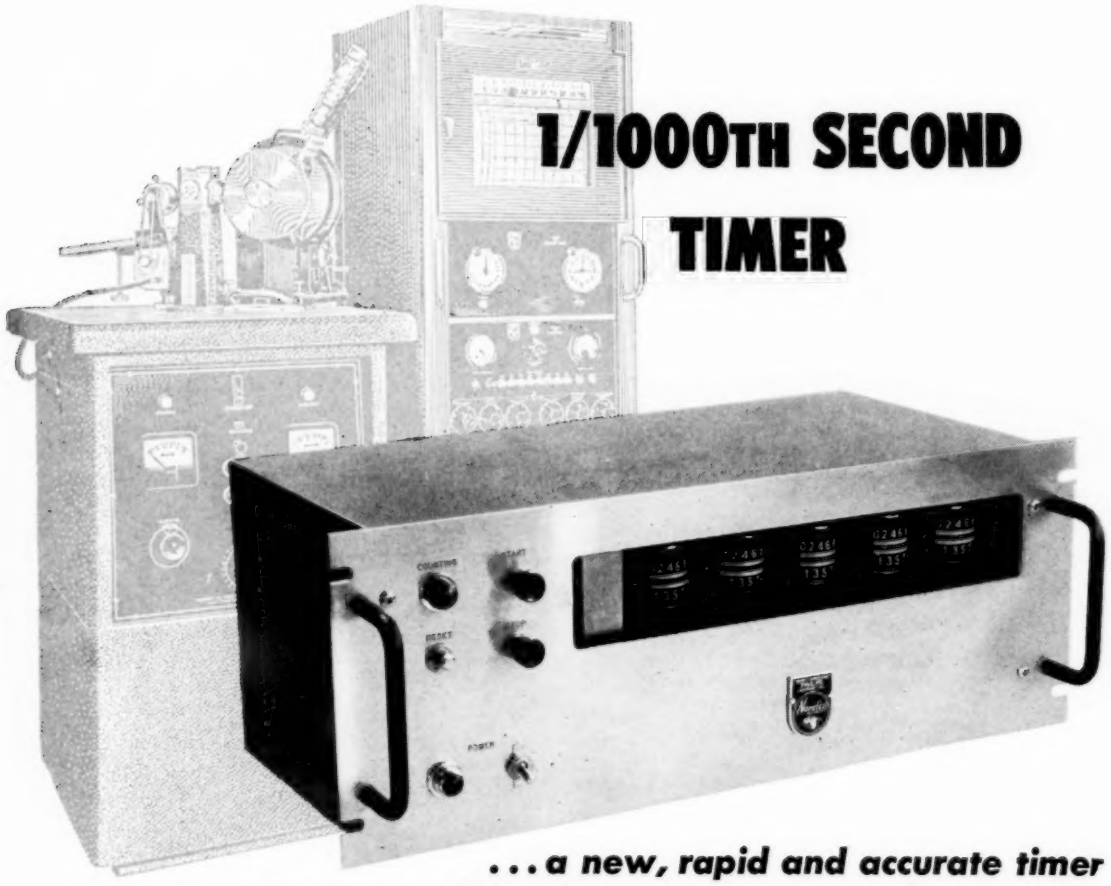
JOURNAL OF APPLIED PHYSICS



1/1000TH SECOND TIMER



...a new, rapid and accurate timer for X-ray diffraction and spectrographic analysis

The new Norelco 100 KC Decade Scaler is a compact direct-reading unit for nuclear measurements and other applications requiring both the high speed and flexibility of accurate electronic counting. Readout is easily observed since the decimals on each of the E1T counter tubes is brightly illuminated. For standard counting, maximium count is 999,999. *(999,999,999 with mechanical stage.)*

Used as a timer to provide accurate counts per second, a tuning fork is wired to the last stage of the unit. The tuning fork oscillator produces a steady 1,000 cycle signal—offering a simple method by which time intervals can be measured to 1/1000th of a second.

The new timer can be installed into the Electronic Circuit Panel and connected by a single lead. To time a fixed count, the timer is set to zero. The goniometer is positioned to the desired 2θ angle and the timer and counting circuits are started simultaneously. When the total number of counts have been registered—both the timer and counting automatically stop. The time is then read directly from the illuminated numbers on the E1T tubes, *in seconds and thousandths of a second.*

With the advent of newly developed Norelco fast detector tubes, *(scintillation and proportional with counting rates over 100,000 cps),* the need for highly accurate timing for measuring small intervals is imperative. At high imput photon rate, the necessary accuracy of a count lasting over a short interval of time, is obvious.

A Norelco PRODUCT NORTH AMERICAN PHILIPS COMPANY, INC.

Serving Science and Industry

750 South Fulton Avenue, Mount Vernon, N. Y.

In Canada: Rogers Majestic Electronics Ltd., 11-19 Brentcliffe Road, Leaside, Toronto 17, Ontario

JOURNAL OF APPLIED PHYSICS

Vol. 26, No. 5 MAY, 1955

In This Issue

		Page
"Gate Modulation" of Electromagnetic Radiation	Kurt Lehovec	495
Note on the Kramers-Kronig Relations	Malcolm K. Brachman	497
Diffusion, Static Charges, and the Conduction of Electricity in Nonmetallic Solids by a Single Charge Carrier. I. Electric Charges in Plastics and Insulating Materials	Selby M. Skinner	498
Diffusion, Static Charges, and the Conduction of Electricity in Nonmetallic Solids by a Single Charge Carrier. II. Solution of the Rectifier Equations for Insulating Layers	Selby M. Skinner	509
Small Particle X-Ray Scattering by Fibers, Size and Shape of Microcrystallites	A. N. J. Heyn	519
Čerenkov Radiation of Electrons Moving Parallel to a Dielectric Boundary	J. G. Linhart	527
Silicon Solar Energy Converters	M. B. Prince	534
Effect of Air Damping on Transverse Vibrations of Stretched Filaments	D. W. Stauff and D. J. Montgomery	540
Plasma Ion Oscillations in Electron Beams	K. G. Hernqvist	544
Anisotropic Diffusion Lengths in Diffusion Theory	Bernard I. Spinrad	548
Asymptotic Spherical Shock Decay	L. M. Tannenwald	551
Propagation of Shock Waves in Aluminum	H. Dean Mallory	555
Noise in One-Dimensional Electron Beams	Hermann A. Haus	560
Arcing at Electrical Contacts on Closure. Part VI. The Anode Mechanism of Extremely Short Arcs	W. S. Boyle and L. H. Germer	571
Multiple Coincidence Magnetic Storage Systems	Robert C. Minnick and Robert L. Ashenhurst	575
Nucleation of Ammonium Iodide Crystals from Aqueous Solutions	J. B. Newkirk and D. Turnbull	579
Self-Propagating Intermittent Discharge	W. S. Boyle	584
Signal-to-Noise Improvement and the Statistics of Track Populations	Nelson Wax	586
Role of Positive Ions in High-Voltage Breakdown in Vacuum	H. C. Bourne, Jr., R. W. Cloud, and J. G. Trump	596
Nyquist's and Thevenin's Theorems Generalized for Nonreciprocal Linear Networks	R. Q. Twiss	599
Impossible Behavior of Nonlinear Networks	R. J. Duffin	603
Cross Correlation between Velocity and Current Fluctuations in Tube Noise	S. V. Yadavalli	605
Variation of Magnetoresistivity of AuCu with Degree of Order	Bernard Wiener, Philip Schwed, and Gerhart Groetzinger	609
Equivalent-Radius Method for Determination of Non-Newtonian Flow Curves from Viscometric Data	K. H. Cram and J. C. Whitwell	613
Resonant-Cavity Field Measurements	S. W. Kitchen and A. D. Schelberg	618
Noise at the Potential Minimum in the High-Frequency Diode	D. A. Watkins	622
Composition of the Interelectrode Prebreakdown Current in High Vacuum	Henry C. Bourne, Jr.	625
Interaction of Moving Charges with Wave Circuits	J. R. Pierce	627
Letters to the Editor:		
Improved "Impregnated Cathode"	Roberto Levi	639
Electron Beam Method of Determining Density Profiles Across Shock Waves in Gases at Low Densities	Douglas Venable and Daniel E. Kaplan	639
Internal Friction of an Alloy of 16 Percent Aluminum in Iron	C. Wert	640
Optical Determination of Base Width in Grown *n-p-n* Silicon Crystals	Alan Fowler and Pascal Levesque	641
Neutron Absorption in a Semi-Infinite Region	J. S. Dwork and H. Hurwitz, Jr.	642
Erratum: On the Interaction of Two Similarly Facing Plane Shock Waves	C. A. Ford and I. I. Glass	643
Research Reports		643
Books Reviewed		643
Books Received		644
Announcements		644
Index to Advertisers		xxvi

JOURNAL OF APPLIED PHYSICS

A journal of applications of physics to other sciences and to industry

Published by the

American Institute of Physics

OFFICERS

FREDERICK SEITZ
Chairman

GEORGE B. PEGRAM
Treasurer

GOVERNING BOARD

ALLEN V. ASTIN
ROBERT F. BACHER
H. A. BETHE
J. W. BUCHTA
S. A. GOUDSMIT
DEANE B. JUDD
HUGH S. KNOWLES
W. H. MARKWOOD, JR.
WILLIAM F. MEGGERS
PHILIP M. MORSE
BRIAN O'BRIEN
HARRY F. OLSON
R. F. PATON
GEORGE B. PEGRAM
ERIC RODGERS
RALPH A. SAWYER
FREDERICK SEITZ
WILLIAM SHOCKLEY
H. D. SMYTH
J. H. VAN VLECK
MARK W. ZEMANSKY

Consultant on Printing

MELVIN LOOS

ADMINISTRATIVE STAFF

HENRY A. BARTON
Director

WALLACE WATERFALL
Executive Secretary

THEODORE VORBURGER
Advertising Manager

RUTH F. BRYANS
Publication Manager

EDITH I. NEFTEL
Circulation Manager

KATHRYN SETZE
Assistant Treasurer

Editor
Robert L. Sproull

Editorial Board

Maurice A. Biot	Elmer Hutchisson	W. R. Smythe
Walker Bleakney	D. O. North	Arthur V. Tobolsky
Wallace J. Eckert	Louis A. Pipes	A. M. Weinberg

THE JOURNAL OF APPLIED PHYSICS is devoted to the application of physics to other sciences, to engineering, and to industry. Previous to 1937, this journal was called PHYSICS.

The JOURNAL OF APPLIED PHYSICS welcomes articles of two distinct kinds: (1) Descriptions of new results in experimental or theoretical physics which are of general interest to applied physicists. The author should explain his contribution in such a way that a physicist or engineer in another field of specialization can appreciate its objectives and essential features. (2) Critical reviews of particular areas of applied physics. Authors contemplating writing such reviews should correspond with the Editor before they begin writing.

The JOURNAL OF APPLIED PHYSICS welcomes letters to the editor which describe in a few words the content of laboratory reports and theses and which tell readers how they may obtain copies of such reports.

Manuscripts should be submitted to the Editor, Rockefeller Hall, Cornell University, Ithaca, New York. Authors should read the "Information for Contributors," facing page 1 of the January issue, before writing articles or letters. The *Style Manual* of the American Institute of Physics is also helpful to writers preparing manuscripts for journals published by the American Institute of Physics. It can be obtained for $1.00 from the American Institute of Physics, 57 East 55th Street, New York 22, New York.

Corrected proofs and all correspondence concerning papers in the process of publication should be addressed to the Publication Manager, American Institute of Physics, 57 East 55 Street, New York 22, New York.

Subscriptions, renewals, and orders for back numbers should be addressed to the American Institute of Physics, 57 East 55 Street, New York 22, New York.

Subscription Price	U.S. and Canada	Elsewhere
To members of the American Institute of Physics and Affiliated Societies	$10.00	$11.00
To all others	12.00	14.00

Back Number Prices

Yearly back number rate when complete year is available: $15.00.
Single copies: $1.50 each.

Changes of address should be sent to the American Institute of Physics.

Advertising rates supplied on request. Orders, advertising copy, and cuts should be sent to the American Institute of Physics.

Entered as second class matter January 22, 1937, at the Post Office at Lancaster, Pennsylvania, under the Act of March 3, 1879. Accepted for mailing at the special rate of postage provided for in paragraph (d-2), section 34.40, P. L. & R. of 1948, authorized May 2, 1932.

Journal
of
Applied Physics

Volume 26, Number 5 May, 1955

"Gate Modulation" of Electromagnetic Radiation

KURT LEHOVEC
Sprague Electric Company, North Adams, Massachusetts
(Received August 16, 1954)

The space-charge layer at a *p-n* junction is a region of low concentration of electrons and holes, and therefore exhibits less absorption than the bulk of the semiconductor for radiation of sufficiently long wavelengths. Thus the space-charge region of a *p-n* junction provides an optical window of a width which can be modulated by an applied field. A quantitative estimate of this effect is given. Because of an obvious relation to unipolar transistors, the name "gate modulation" is proposed, as opposed to the "injection modulation" discussed previously.

INTRODUCTION

IT is well known that the absorption of electromagnetic radiation by a semiconductor in the region at the long wave side of its absorption edge depends on the presence of free carriers (electrons and holes).[1]

It is also well known that the concentration of free carriers can be varied by means of electric fields applied to the semiconductor. These two facts enable the construction of semiconducting devices for modulation of electromagnetic radiation by means of an applied field. One such device utilizes the injection of minority carriers through a rectifying contact[2] or a *p-n* junction[3] when a current is passed in the forward direction. We propose the name "injection modulation" for the change in absorption caused by these minority carriers (and by the change in majority carrier concentration accompanying the injection of minority carriers).

PRINCIPLE OF THE GATE MODULATOR

It is the purpose of this note to describe another type of variation of carrier concentration which may be used for the modulation of radiation. Consider a semiconductor containing a *p-n* junction (Fig. 1). It is well known that there exists a space-charge region around the *p-n* junction, from which free carriers, electrons and holes, are absent. Therefore, this region acts as an optical window (slit or gate) for radiation passing in the direction parallel to the *p-n* junction. It is well known that the width of the space-charge region varies with an applied potential; the "gate" widens with increasing potential applied in the blocking direction. The width of the gate follows the applied voltage up to frequencies limited only by the RC constant of the arrangement, or the relaxation time of the bulk semiconductor. Modulations up to several hundred megacycles per second should be obtainable.

THEORY OF THE GATE MODULATOR

Let g_0 be the incident radiation density (reflection losses already included). Let the absorption constant due to free electrons of concentration n be $1/x_{0,n}=an$, and that due to free holes of concentration p be $1/x_{0,p}=bp$. Let the bulk concentration of electrons in

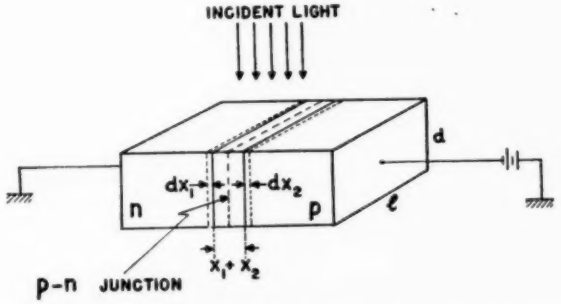


FIG. 1. The space-charge region of width x_1+x_2 acts as an optical "window" for the incident light of sufficiently long wavelengths. The width x_1+x_2 can be varied by the applied potential.

[1] Kaiser, Collins, and Fan, Phys. Rev. **91**, 1380 (1953) for germanium; H. B. Briggs, Phys. Rev. **77**, 727 (1950) for silicon.
[2] A. F. Gibson, Proc. Phys. Soc. (London) **B66**, 588 (1953).
[3] K. Lehovec, Proc. Inst. Radio Engrs. **40**, 1407 (1952); R. Newman, Phys. Rev. **91**, 1311 (1953); H. B. Briggs and R. C. Fletcher, Phys. Rev. **91**, 1342 (1953).

TABLE I. Relations between capacitance C, applied voltage V, transmitted radiation G, and the width of the junction x_1+x_2.[a]

	(1) Step function from N to P	(2) Gradual transition from N to P
(1) C	$\epsilon\epsilon_0 ld/(x_2+x_2)$	$\epsilon\epsilon_0 ld/(x_1+x_2)$
(2) V	$(e/2\epsilon\epsilon_0)(Nx_1^2+Px_2^2)$	$(eq/6\epsilon\epsilon_0)(x_1^3+x_2^3)$
(3) Relation between x_1 and x_2	$Nx_1=Px_2$	$x_1=x_2$
(4) dG/dV	$(g_0C/edNP)$ $\times[P(1-e^{-aNd})$ $+N(1-e^{-bPd})]$	$4g_0C^2/(ld^2qe\epsilon\epsilon_0)$ $\times(2-e^{-aqx_1d}$ $-e^{-bqx_1d})$
(5) dG/dV (if $aNd\ll1$; $bPd\ll1$)	$g_0C(a+b)/e$	$2g_0C(a+b)/e$

[a] The relations of lines (1) to (3) are discussed by W. Shockley, *Electrons and Holes in Semiconductors* (D. Van Nostrand Company, Inc., New York, 1950).

the n-type part be N and that of holes in the p-type part be P. The radiation penetrating a layer of thickness d of the bulk semiconductor is then attenuated by the factor $\exp(-aNd)$ or $\exp(-bPd)$, respectively. In the case of a sudden change from p-type to n-type (step function) the width of the junction x_1+x_2 is related to the junction capacitance C and to the applied voltage V[4] by the well-known equations listed in Table I, column 1. The relations between x_1 and x_2 of line 3 of Table I expresses that negative space-charge on one side and positive space-charge on the other side of the junction compensate each other. The change of transmitted light intensity, G, with increased applied voltage in the blocking direction of the p-n junction arises from the expansion of the space-charge layers by dx_1 and dx_2. The absorption constant in these regions then vanishes, while initially it had the finite value aN and bP, respectively. The change of transmitted light intensity with voltage applied at the junction

$$dG/dV=g_0l[(1-e^{-aNd})(dx_1/dV)+(1-e^{-bPd})(dx_2/dV)]$$

is calculated from the equations of Table I (column 1, lines 1–3) and is listed in lines 4 and 5 of the Table. The last column of Table I contains the corresponding expressions for a gradual junction with a constant gradient q of impurity concentrations.

For germanium, one has[5] $a+b\sim10^{-16}$ cm^2. A typical value of capacitance for a p-n junction of a cross section of 0.5×0.01 cm^2 may be $10^{-11}F$, corresponding to a junction width of 5×10^{-4} cm. Hence, we have $[e/C(a+b)]\sim10^{10}$ volt/cm^2. For $g_0\sim1$ watt/cm^2, we obtain then $(dG/dV)\sim10^{-10}$ watt/volt.

Because of the comparatively narrow width, x_1+x_2 of the space-charge layer, the junction would have to be rather accurately plane across the thickness d of the sample.

At a width of the p-n junction of the order indicated above, diffraction phenomena may occur for infrared light. Dr. Ray Jackson of this laboratory has proposed to use a grating of p-n junctions (such as may be prepared by the method of Hall[6]) for focusing of the transmitted light in certain directions.

RELATION BETWEEN RADIATION MODULATION AND TRANSISTORS

In conclusion, I would like to point out an interesting analogy between (a) unipolar transistors[7] and gate modulators, and (b) bipolar transistors and injection modulators. The devices of group (a) depend on the change of width of a gate by the widening of a space-charge layer; they are high input-impedance devices with extremely short time-constants of response. The devices of group (b) depend on injection of minority carriers; they are low input-impedance devices of a time-constant determined essentially by the lifetime of minority carriers or by transit considerations.

Acknowledgment is made to Miss A. Marcus, Mr. K. Schoeni, and Dr. R. Jackson for helpful discussions.

[4] Strictly, applied voltage plus diffusion potential; the latter is almost negligible at an applied voltage of several volts in the blocking direction.

[5] Kaiser, Collins, and Fan, Phys. Rev. **91**, 1380 (1953).
[6] R. N. Hall, Phys. Rev. **88**, 139 (1952).
[7] W. Shockley, Proc. Inst. Radio Engrs. **40**, 1365 (1952).

Note on the Kramers-Kronig Relations

MALCOLM K. BRACHMAN
Texas Instruments, Incorporated, Dallas 9, Texas
(Received July 29, 1954)

The explicit form of the relationship between the real and imaginary parts of a physically realizable network function is correlated with the alternate conventions available for such problems. The transformation from a linear frequency scale to a logarithmic one is simply effected by considering the Mellin transforms of the appropriate functions.

THE relationship between the real and imaginary parts of a physically realizable network function has been exploited in many diverse applications.[1] For our purposes the term network function shall subsume driving point and transfer immittances, complex susceptibilities (electric or magnetic), scattering matrices, and a number of similar quantities. The network function $Q(\omega)$ is a function of the complex frequency ω. A simple example is the impedance of the parallel combination of a resistor R and a capacitor C, which is

$$Q(\omega)=\frac{R(1-i\omega CR)}{1+(\omega RC)^2}. \tag{1}$$

It may be noted that some discretion is required in choosing the network function. For instance, the choice of the admittance rather than the impedance for the parallel R-C combination leads to divergent integrals.

The fact that $Q(\omega)$ is an analytic function of ω assures us that there will be some relationship between its real and imaginary parts. A derivation of the relationship from this point of view is given in Chapter VI of Guillemin's book,[2] where the equations we call the Kramers-Kronig relations are Eqs. (369) and (370). Bode[3] presents a different treatment in Chapters XIII and XIV of his book. The appropriate equations, shown in Chapter XIV, are (3) and (21). A cursory survey of the literature indicates that the explicit form of these relationships varies from author to author. It seems worthwhile to provide an understanding of the differences, which stem from different conventions.

We assume a network function $Q(\omega)$ obeying the Kramers-Kronig relations. We may write either

$$Q(\omega)=J(\omega)-iH(\omega) \tag{2a}$$

or

$$Q(\omega)=A(\omega)+iB(\omega). \tag{2b}$$

We may now take the admittance of a capacitance in the alternate forms

$$Y=i\omega C \tag{3a}$$

and

$$Y=-i\omega C, \tag{3b}$$

corresponding to the time factors $e^{i\omega t}$ and $e^{-i\omega t}$. We write the Kramers-Kronig relations in the form appropriate for the combination (2a) and (3a):

$$J(\omega)=\frac{2}{\pi}\int_0^\infty \frac{H(y)ydy}{y^2-\omega^2} \tag{4}$$

and

$$H(\omega)=\frac{2\omega}{\pi}\int_0^\infty \frac{J(y)dy}{\omega^2-y^2}. \tag{5}$$

These integrals are to be interpreted as Cauchy principal values.

Table I now gives the quantities to be substituted for J and H in each of the four cases.

We have followed convention (2a), (3a) in our previous work and will do so in the remainder of this paper.

It has been shown that the Mellin transforms[4] $j(s)$ and $h(s)$ of $J(\omega)$ and $H(\omega)$ satisfy the relation[1]

$$h(s)=j(s)\tan(\pi/2)s. \tag{6}$$

A particular pair obeying this equation are

$$h_0(s)=(1/s)\tan(\pi/2)s \tag{7a}$$

and

$$j_0(s)=1/s. \tag{7b}$$

The inverse Mellin transforms of this pair are

$$H_0(\omega)=\frac{1}{\pi}\ln\left|\frac{1+\omega}{1-\omega}\right| \tag{8a}$$

and

$$\begin{aligned} J_0(\omega)&=1, \quad \omega<1 \\ &=0, \quad \omega>1. \end{aligned} \tag{8b}$$

TABLE I.

	J	H	Authors
Case (2a), (3a)	J	H	Pake and Purcell[a]
Case (2b), (3a)	A	$-B$	Bode, Guillemin
Case (2a), (3b)	J	$-H$	
Case (2b), (3b)	A	B	Fröhlich[b]

[1] M. K. Brachman and J. R. Macdonald, Physica **XX**, 1266 (1954).
[2] E. A. Guillemin, *The Mathematics of Circuit Analysis* (John Wiley and Sons, Inc., New York, 1949).
[3] H. W. Bode, *Network Analysis and Feedback Amplifier Design* (D. Van Nostrand Company, Inc., New York, 1945).
[a] G. E. Pake and E. M. Purcell, Phys. Rev. **74**, 1184 (1948).
[b] H. Fröhlich, *Theory of Dielectrics* (Clarendon Press, Oxford, England, 1949).
[4] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals* (Clarendon Press, Oxford, England, 1937).

We may transform (4) to an important equivalent form by making use of these relations.

For this purpose we require two further facts. The first is that the inverse Mellin transform of $f(s)g(s)$ is $\int_0^\infty F(x)G(\omega/x)dx/x$. The second is that the inverse Mellin transform of $sj(s)$ is $-\omega dJ/d\omega$, if proper assumptions are made about the behavior of $J(\omega)$ at zero and infinite frequency. We may now factor (6) into the form

$$h(s)=[sj(s)]\left[\frac{1}{s}\tan\frac{\pi}{2}s\right], \tag{9}$$

and it follows from the above theorems concerning inverse Mellin transforms that

$$H(\omega)=-\frac{1}{\pi}\int_0^\infty x\frac{dJ}{dx}\ln\left|\frac{1+\frac{\omega}{x}}{1-\frac{\omega}{x}}\right|\frac{dx}{x}$$

$$=-\frac{1}{\pi}\int_0^\infty \frac{dJ}{dx}\ln\left|\frac{x+\omega}{x-\omega}\right|dx. \tag{10}$$

This equation is the same as Bode's Eq. (17) as shown in Chapter XIV.[3] The difference in sign is due to the difference in conventions.

From Eq. (6) we may deduce in the same manner that

$$j(s)=[(1-s)h(s)]\left[\frac{1}{1-s}\tan\frac{\pi}{2}(1-s)\right]. \tag{11}$$

Using the fact that the inverse Mellin transform of $(1-s)h(s)$ is $d/d\omega\{\omega H(\omega)\}$ and following the procedure of reference 1, we obtain

$$J(\omega)=\frac{1}{\pi\omega}\int_0^\infty \frac{d}{dx}[xH(x)]\ln\left|\frac{\omega+x}{\omega-x}\right|dx. \tag{12}$$

This corresponds to Bode's Eq. (25), Chapter XIV,[3] (with the choice $z=\omega$) except for the expected change in sign.

I am indebted to Dr. J. Ross Macdonald for pointing out the problem of sign conventions in this connection and for interesting discussions on this and related matters.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Diffusion, Static Charges, and the Conduction of Electricity in Nonmetallic Solids by a Single Charge Carrier. I. Electric Charges in Plastics and Insulating Materials*†

SELBY M. SKINNER
Department of Chemistry and Chemical Engineering, Case Institute of Technology, Cleveland, Ohio

(Received June 15, 1954)

The dynamic equilibrium of space charge in insulators such as ceramics, molecular crystals, or high polymers, is considered from the thermodynamic point of view with quantum-mechanical boundary conditions. The treatment constitutes the opposite extreme from the usual ohmic conduction equations; in the first paper, only the zero-current equilibrium of charge is considered. Applications are made to the observed charging of high polymers during molding.

I. INTRODUCTION

IT is well known that high polymers such as polystyrene or polyethylene acquire electrostatic charges in the molding process, and that these charges persist for a long time, even years. In the textile industry, such charges accumulating on synthetic fibers introduce undesirable problems in processing, and therefore, antistatic treatments such as coatings are in common use. They reduce the tendency, for example, of plastic wall tile or of synthetic fiber shirts to accumulate dust and improve the behavior of synthetics in the handling stages required during manufacture. In certain cases, such as in the "operating theater" of a hospital, it is desirable to reduce as far as possible the chances of electrostatic charge buildup and subsequent sparking.

Woodland and Ziegler[1] have shown that the observed charge is greater the greater the resistivity of the high polymer, that no charge is obtained on the parts which did not touch metal during the molding process, that excessive lubrication during molding reduces the charge observed, that plastic flow during molding has only slight (if any) effect, and that no destaticizer tried remains efficient after washing. In recent work,[2] it has

* Initially received November 3, 1953. Received in final form June 15, 1954. Portions of this and the next article were presented at the Detroit meeting of the American Physical Society, March, 1954, and at two Gordon Research Conferences on adhesion, June 28–July 2, 1954, and on organic coatings, July 12–16, 1954.
† With the support of the Aeronautical Research Laboratory, Wright Air Development Center, U. S. Air Force.

[1] P. C. Woodland and E. E. Ziegler, Modern Plastics **28**, 95 (May, 1951).
[2] Skinner, Savage, and Rutzler, Jr., J. Appl. Phys. **24**, 438 (1953); **25**, 1055 (1954). The present paper incorporates certain material in correspondence to Messrs. West and Morant which could not be included in the letter because of space limitations.

been shown that upon the break of an adhesive bond between high polymer and metal, an electrical potential is obtained; this is indicative of a charge transfer between the two prior to the break. Efforts to permanently neutralize this charge after break, by surface treatment, only, of the high polymer, have been unsuccessful, and it must be concluded that it is a volume charge distribution within the polymer. In the industry, it has also been found that surface treatment, only, is insufficient to remove such charges from plastic materials,[3] i.e., that they have the characteristics of volume charges.

Several explanations of such a volume charge distribution are possible, but the results of Woodland and Ziegler appear to favor the mechanism further discussed in the article. Such a mechanism is in direct contradiction to the point of view[4] that charges on insulators are true surface charges and can be explained entirely in terms of surface states. However, as will be seen in the present and the following article, not only does the assumption of a volume charge distribution put into proper order and explain observed volume effects, but also in the appropriate case the charge, though still a volume distribution, decreases so rapidly from the surface of the insulator inwards, that experimentally it would be difficult to make a distinction between the two.

1. The Properties of an Insulator

In gases and solutions, space charge does not exist to any large degree in the equilibrium state, since any disparity in charge concentration is soon corrected by motion of the charged particles. In solid insulating materials, on the other hand, due to the high resistivity, correction of such discrepancies would take a very long time; if any mechanism exists which favors the formation of a space charge, an equilibrium distribution of electricity should be characterized by a volume distribution of charge throughout the insulator.

High polymers and ceramics have been under commercial manufacture for the electrical industry for many years, and a continually improved product has resulted. Major requirements in electrical usage have been high electrical resistance, low loss factor (energy dissipation), and high electrical breakdown strength. To achieve a low loss factor, it is desirable that the material be nonpolar, and have a minimum of phase boundaries or internal cracks; polarity would increase the Debye absorption, and internal nonhomogeneity would increase the Maxwell-Wagner loss. High surface and volume resistance imply a large forbidden energy gap between the valence and conduction levels, and freedom from imperfections which would yield conduction paths

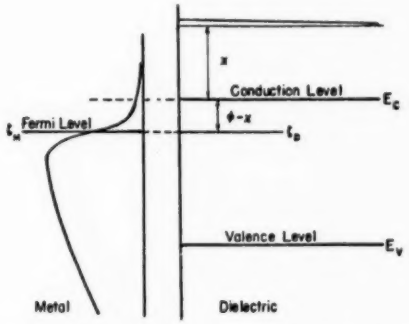


FIG. 1. Energy levels and charge distribution in metal-insulator contact.

and from impurities which would give energy levels in the forbidden zone. High breakdown strengths imply few naturally free charges and few cracks within the material (which by the v. Hippel or the Fröhlich mechanisms would permit electrons to achieve the energy necessary to initiate breakdown).

A well-manufactured ceramic or high polymer such as for electrical use, may therefore be thought of as a homogeneous material without donor or acceptor levels, with a high forbidden energy band of the order of from 2–7 ev between the valence levels and the conduction levels.

The observations of Woodland and Ziegler that no charge is found on that portion of the polymer which has not touched metal, and those in reference 2 suggest that the charge in the insulator comes from the Fermi distribution in the metal.[5] It is well known that the contact potential between two different metals is because of the alteration of the electron concentration in each by charge transfer and is a consequence of the equalization of the Fermi levels. In the contact between a metal and an insulator, equilibrium is also not reached until the Fermi levels are equal on both sides of the interface. Since the (negative) charges cannot travel into the insulator in the valence level of energies, equalization can only occur by transfer into or out from the conduction levels of the insulator, as shown in Fig. 1. Such charge transfer between neutral substances in contact is not predicted by classical electrostatics. The mechanism is diffusion into the insulator, and when equalization has occurred, a dynamic equilibrium exists between the charges in the metal and those in the insulator; accompanying the equilibrium are the phenomena of transferred charge between metal and insulator, volume space charge, increasing height of the conduction level in the interior of the polymer, and an electrostatic attraction between the metal and the polymer discussed in reference 2. The question must then be considered; do instances exist in which the energy gap between the Fermi level of the metal and the conduction band of the insulator is small or even

[3] Woodland and Ziegler, reference 1, and personal communication from R. F. Boyer, Dow Chemical Company, Midland, Michigan. Similar difficulty is experienced in the radio parts industry.

[4] For example, F. A. Vick, "Theory of contact electrification," a paper in the symposium on Static Electrification, of the Institute of Physics in London, 25–27 March 1953, Supplement No. 2 to the British J. Appl. Phys.

[5] A particular case was treated by N. E. Mott and R. W. Gurney, *Electronic Processes in Ionic Crystals* (Oxford University Press, New York, 1940), pp. 168–173.

negative, so that appreciable charge transfer occurs? Mott and Gurney[5] give examples in inorganic crystals to show that this can be so; the phenomena referred to at the beginning of this article indicate that it can be so for organic materials; and support for this point of view is furnished by observations on the semiconductivity of organic materials[6] and the phenomena mentioned in a number of the papers in the symposium, see reference 4.

2. Contact between a Metal and an Insulator

Since the free charge density in the metal is very much greater than can be transferred into the insulator, the Fermi level in the metal is not appreciably affected by the charge transfer; for the same reason, classical statistics are appropriate in the insulator. Thermodynamic equilibrium requires that the rate of charge transfer be the same at the boundary in both directions. In the metal this is the charge density in the Fermi level times the Boltzmann factor for the energy gap $\varphi-\chi$. Thus, the volume charge density *at the boundary* of the dielectric is given by

$$n_0=2(2\pi mkTh^{-2})^{\frac{3}{2}}\exp[-(\varphi-\chi)/kT]. \quad (1)$$

The depth, χ, of the conduction level in the insulator is the experimentally determinable electron affinity.

Since the equilibrium is dynamic, the free energy will be the same throughout all the phases in contact. Zero net current equilibrium is assumed; the case in which a voltage is applied is taken up in the next paper. Existence of space charge within the nonmetal is integral to the assumptions herein; the way of thinking which follows from considerable familiarity with the field of semiconductors or electrolytic solutions can give rise to inconsistencies if this is not kept in mind, since in semiconductors for example, when electrons are released to the conduction level, or injected, compensating charges exist in the material either as holes or fixed ionized impurities, or are simultaneously injected. The non-ohmic behavior in nonmetals further discussed in this article follows directly from the fact that space charge and conduction by a single type of uncompensated carrier is the opposite extreme of the electrical flow equations usually considered. Since the equilibrium is dynamic, and the charge concentration is not uniform in the space within the insulator, the diffusion equation is applicable.

II. THE DIFFUSION EQUATION

1. A Single Type of Charge Carrier

The vector equation for the current owing to a single type of charge carrier (e.g., electron or hole), when account is taken of diffusion, is

$$\vec{j}=nebF-De\nabla n. \quad (2)$$

Here $\vec{j}$ is the current per cm^2 of surface normal to the current, e the electric charge of the carrier (positive for holes), n the volume density of charge carriers, D the diffusion constant, b the mobility of the carrier in the material,[7] $\bar{F}$ the electric field strength, and ∇ denotes the gradient.

If there is no diffusion, if the density n is independent of the space coordinate (as in first approximation in a metal), or if the electric field, charge density, or mobility, is so great that the second term is negligible with respect to the first, only the first term need be retained, the usual expressions

$$\vec{j}=\sigma F, \quad \sigma=neb, \quad (2a)$$

result, and Ohm's law follows when the charge density is constant.

If, on the other hand, the diffusion term is included, it is not to be expected in general that relations as simple as those resulting from Eq. (2a) will be obtained. In what follows only steady state, constant temperature, conditions are considered. Permanent sources or sinks of charge or current will be assumed to exist only at the boundaries, and therefore, in the interior of the material the equation of continuity for the current j is applicable. The physical situation will be discussed in terms of a one-dimensional space dependence, hence the current in the x direction will be constant in the interior of the material, that in either of the other two directions having at any point a net value of zero. The quantities b and D are assumed independent of the space coordinate (isotropic and homogeneous material) except that changes may occur at boundaries; they are, however, dependent upon the temperature and other parameters. In the one-dimensional case,

$$j=nebF-De(dn/dx). \quad (2b)$$

When diffusion is present, the effect is to produce a current away from regions of high concentration of charge density.

2. Charge Carriers of More than One Kind

With a small forbidden energy gap, or at higher temperatures, appreciable charge from the valence level will enter the conduction level, and charge carriers of both signs must be considered.[8] In this case,

$$\vec{j}_+=b_+n_+e\bar{F}-eD_+\nabla n_+; \quad \vec{j}_-=b_-n_-e\bar{F}+eD_-\nabla n_-; \quad (3.1)$$

$$\vec{j}=\vec{j}_++\vec{j}_-=(b_+n_++b_-n_-)e\bar{F} -e(D_+\nabla n_+-D_-\nabla n_-). \quad (3.2)$$

[6] Eley, Parfitt, Perry, and Taysum, Trans. Faraday Soc. **49**, 79 (1953).

[7] The letter b is used for the mobility to avoid confusion in later equations, since the symbol, μ, usually used for this quantity is also universally used for the chemical potential.

[8] The general case *with, however, the assumption of electrical neutrality and fixed ionic charges* has been treated by a number of authors including: W. Shockley, Bell System Tech. J. **28**, 435 (1949); W. van Roosbroeck, Bell System Tech. J. **29**, 560 (1950); and R. C. Prim, III, Bell System Tech. J. **30**, 1174 (1951). C. Herring, Bell System Tech. J. **28**, 401 (1949), has considered some more general cases without fixed ionic charges, but assuming electrical neutrality and eliminating the use of Poisson's equation; diffusion is neglected until near the end, when an estimate of its effects is given.

To these relations must be added Poisson's equation

$$\nabla\cdot\bar{F}=\frac{4\pi e}{\epsilon}(n_+-n_-). \tag{3.3}$$

If the last term in parentheses in (3.2) is zero, ohmic conduction follows. For equal b's and equal D's, this can occur either by equal diffusion of charges of each sign, or by charge distribution such that the gradient of each is zero; by (3.3) if $n_+=n_-$ everywhere, the field is constant and the diffusion term in (3.2) vanishes, so that Ohm's law is valid. This means that the diffusion of the positives negates the effect of the diffusion of the negatives, and the constancy of current then requires that the concentration of charges be independent of x.

When n_+ differs slightly from n_-, Eqs. (3.2) and (3.3) yield

$$D\frac{d^2F}{dx^2}-bF\frac{dF}{dx}(\pm 1+\delta)+\frac{4\pi j}{\epsilon}=0, \tag{4}$$

where

$$\delta=\frac{2n_{\mp}}{n_+-n_-},$$

and both upper signs are to be taken if $n_+>n_-$, both lower signs if $n_->n_+$. If either n_+ or n_- is small compared to the other $\delta\ll 1$, *a preponderance of charge of one sign makes it necessary to consider diffusion.*

If the current is carried *entirely* by carriers of one sign, δ actually is zero; this is a limiting case just as the case $n_+=n_-$ or $\nabla n_i=0$ (ohmic conduction) is itself a limiting case in the other direction. If there is merely such a preponderance of charges of one sign that δ is small but not identically zero, observed results will be closely represented by the solutions of Eq. (4) for $\delta=0$ but with an adjusted value of b.[9]

Thus for positive carriers,

$$D\frac{d^2F}{dx^2}-bF\frac{dF}{dx}+\frac{4\pi j}{\epsilon}=0, \tag{5}$$

and for negatives the same equation holds with the sign of b changed. It is the basic equation for the treatment that follows; since the Sommerfeld and Bethe treatment[10] of the current through a solid on the basis of the departure from thermodynamic zero current distribution functions for charged particles also yields Eq. (2b) at constant temperature and under classical statistics, Eq. (5) can be accepted as satisfying thermodynamic requirements.

The usual solution of (5) has been obtained[11] by assuming a charge density distribution and calculating other quantities therefrom. While, especially in the cases of small space charge this is frequently a good approximation, it is more reasonable that in the steady state flow of carriers here considered the current should be constant (by the equation of continuity), except at certain types of boundaries. Unless an extremely fortunate choice of charge density distribution has been made, such constancy of current is only approximate. With a large density of space charge, therefore, an exact solution of Eq. (5) must be used. The results should also furnish some indications as to what the effects of space charges will be in other nonmetallic solids which show non-ohmic conduction.

3. Properties of the Diffusion Equation

The Einstein relation connecting the mobility and the diffusion constant

$$eD=bkT \tag{6}$$

has been shown[12] to be valid not only for the static equilibrium case but also for the steady flow of a current pulse through a semiconductor. It is used to replace the diffusion constant by the temperature of the carrier charges.

The equation to be solved is therefore

$$F''-\frac{e}{kT}FF'+\frac{4\pi ej}{\epsilon bkT}=0, \tag{7}$$

and, from the solution,

$$\psi_2-\psi_1=e\int_1^2 F dx;\quad dF/dx=4\pi ne/\epsilon, \tag{8}$$

where ψ is the inner electrostatic potential energy, i.e., the charge on the carrier times the inner electrostatic potential (see Sec. III).

A first integral of Eq. (7) is (see reference 5, page 172)

$$\frac{e}{2kT}\frac{dF}{dx}-\frac{e^2}{4k^2T^2}F^2+\frac{2\pi e^2 j(x+x_0)}{\epsilon bk^2T^2}=g^2. \tag{9}$$

All or any part of the jx_0 term could be included in the g^2; reasons for not doing so will become evident in the current-carrying case. The constants g and x_0 may, of course, depend upon the current, j, or any other parameter. Since the equation is nonlinear, solutions are not

[9] If δ is small but not zero, and Eq. (4) is linearized to the first order in δ, i.e. $F=F_{\delta=0}+F_1$ and only terms in δ retained, the assumption that d^2F_1/dx^2 is small yields $F_1=+F_0/2$. Therefore, a better approximation to the effect of charges thrown up from the valence level may be obtained from the solution for $\delta=0$, that for $\delta\neq 0$ being obtained as $F=(1+\delta/2)F_0$ (4.2) in which δ [Eq. (4)] may be given its local value.

[10] A. Sommerfeld and H. Bethe, *Handbuch der Physik* (Springer-Verlag, Berlin, 1933), second edition Vol. 24.2, p. 333. The treatment is also in R. H. Fowler, *Statistical Mechanics* (The Cambridge University Press, New York, 1936), second edition; and J. E. Mayer and M. G. Mayer, *Statistical Mechanics* (John Wiley and Sons, Inc., New York, 1940).

[11] For example, P. T. Landsberg, Proc. Roy. Soc. (London) **A213**, 226 (1953) who has used the diffusion equations effectively, to investigate the Einstein "mobility-diffusion constant" relation and the rectifying properties of semiconducting materials.

[12] Transistors Teachers Summer School, Phys. Rev. **88**, 1368 (1952); Landsberg, reference 11.

additive, but must be determined for the specific case under consideration.

4. Static Charges: The Zero-Current Equilibrium State

The basic solutions for zero current have been considered by v. Laue[13] and others, but without a clear distinction between inner potentials and Volta potentials. These treatments may be consulted for mathematical details. These cannot be used as given in the literature, since thermodynamic requirements must be incorporated; also, the solution for the current-carrying case depends upon the properties of the various possible physical conditions in the zero-current equilibrium. Accordingly, a brief sketch of the general properties of the solutions[13] is given hereafter, and these are then applied to the practical case of charged insulators.

The three basic solutions differ in form depending upon whether g^2 is positive, negative, or zero. The first one is usually referred to as the finite case, since it is valid only for a finite thickness, nevertheless, as will be seen below, finite portions of the other cases are valid for finite dielectrics, the choice depending upon the boundary conditions, particularly, upon the energy gap at the metal-dielectric boundary. Even the "finite" solution is not valid, *in general*, throughout the full length of the material, but the two solutions must be fitted together by proper boundary conditions.

a. The Case: $g^2>0$, $j=0$. Letting $u=g(x+x_0)$.

$$\left.\begin{aligned}\psi&=2kT\ln\sin u+\text{const}\\ F&=-[2kT/e(x+x_0)]u\cot u\\ n&=[\epsilon kT/2\pi e^2(x+x_0)^2]u^2\csc^2 u\end{aligned}\right\}\tag{10.1}$$

at $x=0$:

$$\left.\begin{aligned}F_0&=-(2kT/e)g\cot gx_0\\ n_0&=(\epsilon kT/2\pi e^2)g^2\csc^2 gx_0\end{aligned}\right\}.\tag{10.2}$$

The quantities are not significant outside the range $0<u<\pi$; they are symmetrical with respect to $u=\pi/2$. At $u=\pi/2$:

$$x=g^{-1}\left(\frac{\pi}{2}-x_0\right);\quad F_1=0;\quad n_1=(\epsilon kT/2\pi e^2)g^2.\tag{10.3}$$

The quantity g^2 is by Eq. (9) given by

$$\frac{d\varphi}{dx}-\varphi^2=g^2>0,\tag{10.4}$$

where $\varphi=eF/2kT$. At the boundary, $x=0$, so that

$$\left.\begin{aligned}n_0&=(\epsilon kT/2\pi e^2x_0^2)\Theta^2\\ \text{and}\qquad\\ x_0&=(\epsilon kT/2\pi e^2n_0)^{\frac{1}{2}}\Theta\end{aligned}\right\}\tag{10.5}$$

analogous to the characteristic length in the Debye-Hückel theory. It ranges from the astronomical to the submicroscopic, decreasing as the volume charge density at the boundary increases. Values [for solution (11), following] are given in reference 2; the expression, Θ, may be regarded as a correction factor to x_0 imposed by the finite dimension of the dielectric. This, with a change of origin, is the case used by Fowler[13] in his first discussion of electron atmospheres, and is useful with a finite dielectric between two identical metal surfaces. If g^2 is large, the solution refers to a thin layer of material; however, at sufficiently small thicknesses, less than those encountered in practice, the fundamental assumptions are not valid, because of tunneling and insufficient collisions for diffusion equilibrium. In Eqs. (10) as compared with (11) and (12), the inner electric field is small in comparison with the charge density. In fact,[14] φ is a measure of the ratio of the distance traveled by the charged particle because of the electric field to the distance it would travel under diffusion in the same time.[15]

b. The Case: $g^2=0$, $j=0$.

$$\left.\begin{aligned}\psi&=2kT\ln(1+x/x_0)+\text{const}\\ F&=-(2kT/e)(x+x_0)^{-1}\\ n&=(\epsilon kT/2\pi e^2)(x+x_0)^{-2}\end{aligned}\right\}\tag{11.1}$$

at $x=0$:

$$F_0=-2kT/ex_0;\quad n_0=\epsilon kT/2\pi e^2x_0;\tag{11.2}$$

at $x=\infty$:

$$\psi_\infty=0[\ln x];\quad F_\infty=n_\infty=0;\tag{11.3}$$

and

$$d\varphi/dx-\varphi^2=0.\tag{11.4}$$

This corresponds to the simple case of an infinite solid with no field at infinity. To a very high degree of accuracy, it may be used as an approximation to the other solution in many practical cases, as may be seen

[13] M. v. Laue, Jahrb. Radioakt. u. Elektronik **15**, pp. 205, 257, 301 (1928); Handb. d. Radiol. **6**, 452 (1925). W. Schottky, Jahrb. Radioakt. u. Elektronik **12**, 14 (1915). O. W. Richardson, *The Emission of Electricity from Hot Bodies* (Longmans Green and Company, New York, 1921), second edition. R. H. Fowler, *Statistical Mechanics* (The Cambridge University Press, New York, 1936), second edition. F. Borgnis, Z. Physik **100**, 117 (1936). After this discussion was in manuscript form, a letter was also received from D. C. West of the Nova Scotia Research Foundation in which the existence of the three solutions was pointed out independently. Dr. West's letter is in J. Appl. Phys. **25**, 1054 (1954).

[14] W. Shockley, *Electrons and Holes in Semiconductors* (D. van Nostrand Company, Inc., New York, 1950), p. 350; J. R. Haynes and W. Shockley, Phys. Rev. **81**, 835 (1951).

[15] A rough conceptual picture can also be obtained as follows: let δ be the distance a particle of energy kT could go against the electric field before being brought to rest, reversed in direction, and given again an energy kT, but now traveling in the opposite direction. Then

$$\delta eF=-2kT;\quad \frac{d\varphi}{dx}-\varphi^2=\delta^{-2}\left(\frac{d\delta}{dx}-1\right)$$

and if $g^2\gtrless 0$, $d\delta/dx\gtrless 1$, which relates the change of this characteristic distance, δ, to the change in the space coordinate. Or again, since $2kT/e$ is the ratio, (by Sommerfeld and Bethe) of the transport of heat, or (by irreversible thermodynamics) of the transport of the energy distinct from heat, to the electric current under classical isothermal conditions, φ is related to the ratio of the energy due to the electric potential to that transported by the transfer of electrons.

by expanding the circular and hyperbolic functions in terms of u or v. The necessary condition is that x_0 be small, i.e., that n_0 be large. The solution is also applicable to a finite thickness of dielectric when the boundary conditions imply electric fields and charge densities at the interfaces, of the same value as would be found in an infinitely thick dielectric if sectioned at the appropriate values of x.

c. The Case: $g^2<0$, $j=0$. Let $-g^2\equiv a^2$, $v=a(x+x_0)$.

$$\left.\begin{aligned}\psi&=2kT\ \ln\sinh v+\text{const}\\ F&=-[2kT/e(x+x_0)]v\ \operatorname{ctnh}v\\ n&=[\epsilon kT/2\pi e^2(x+x_0)^2]v^2\ \operatorname{csch}^2 v\end{aligned}\right\}\qquad(12.1)$$

at $x=0$:

$$\left.\begin{aligned}F_0&=-(2kT/e)a\ \operatorname{ctnh}ax_0;\\ n_0&=(\epsilon kT/2\pi e^2)a^2\ \operatorname{csch}^2 ax_0;\end{aligned}\right\}\qquad(12.2)$$

at $x=\infty$:

$$\psi_\infty=0[x];\quad F_\infty=-2kTa/e;\quad n_\infty=0.\qquad(12.3)$$

As before, the second equation in (12.2) permits determining x_0, and the equation for determining a^2 is

$$d\varphi/dx-\varphi^2=-a^2<0.\qquad(12.4)$$

This case arises when the effect of the internal electric field is greater than that of the charge density. An electric field throughout the sample, directed toward the origin, accompanied by no current, may seem a contradiction. It is due to the rapid decrease of charge density, so that the diffusion current is always great enough to balance that due to the field.

For Eqs. (12) to be a solution, u must be positive, hence at infinity the field is directed towards the origin. For the same reason, the charge density is always positive. For positive a, the solution is valid in the range $-x_0<x<\infty$ but gives infinite values at $x=-x_0$; it represents the electrical equilibrium to the right of an infinite sheet of charge at $-x_0$; the actual boundary is mathematically replaced by a continuation of the dielectric to such a charge sheet, and diffusion proceeds from it into the actual dielectric. As the charge density on the actual boundary decreases, the diffusing plane moves back from the boundary. When "a" (not a^2) is negative, the mirror image condition is obtained.

III. FERMI LEVEL OR ELECTROCHEMICAL POTENTIAL

1. Inner Potential

The electrostatic potential energy, ψ, just determined is an inner potential times the charge of the carrier. This has not always been brought out in the derivations of the solutions, and differences of the inner potential have, at times, been used as equivalent to differences of Volta potentials. A consideration of the over-all thermodynamic picture shows that in the case of a single charge carrier, in excess, the inner electrostatic[16] potential may be defined in terms of χ, i.e., the depth of the conduction level below the potential of the electron vapor external to the material. Taking the Fermi level as the zero of electrostatic potential, the height of the conduction level differs from the inner potential only by an arbitrary constant *which is the same for all materials in contact*, and which may be taken as zero. Thus the well-known lack of significance of the Galvani potential as compared to the electrochemical potential within the material does not enter into question. The electrochemical potential (Fermi level) is given by

$$U=\psi+\mu=\psi+kT\ \ln n-c\qquad(13)$$

with the same symbols as before and μ the chemical potential; therefore,

$$\psi=U-kT\ \ln n-c.$$

In the space charge picture, U corresponds to Shockley's[17] quasi-Fermi level, (imref); the imref as shown elsewhere[16] performs the important function of substituting the correct (electrochemical) potential for the inner electrostatic potential in the discussion of current flow.

2. Volta Potential

In the current flow through solids, the external voltage, V, between metal electrodes at either end of the solid, is related (see reference 17, pp. 464–5) to the electrochemical potentials of its two ends by

$$eV=U_2-U_1.\qquad(14)$$

Defining the electrochemical potential by Eq. (13) for the appropriate values of x, it is readily seen that the solutions in Eqs. (10) to (12), although exhibiting inner potentials, correspond to an electrochemical potential which is independent of the spatial coordinate. Thus, its gradient is zero.

Either by the diffusion Eqs. (3.1) by the Sommerfeld-Bethe treatment of current flow,[10] or by the thermodynamics of the steady state[18] it is possible to show that the generalized current (electrical current in this case) is

$$\vec{j}=-nb\nabla U=-nb\frac{dU}{dx}.\qquad(15)$$

Thus the solutions in Eqs. (10) to (12) correspond

[16] S. M. Skinner (to be published).

[17] W. Shockley, Bell System Tech. J. **28**, 435 (1949); also, *Electrons and Holes in Semiconductors* (D. Van Nostrand Company, Inc., New York, 1950).

[18] S. R. DeGroot, *The Thermodynamics of Irreversible Processes* (North Holland Publishing Company, Amsterdam), and (Interscience Publishers, Inc., New York, 1951). K. G. Denbigh, *The Thermodynamics of the Steady State* (Methuen and Company, Ltd., London), and (John Wiley and Sons, Inc., New York, 1951). The special case is diffusion and electrical conduction, U is the generalized potential, $X(=-\nabla U=J/L)$ the generalized force, and $L(=nb)$ the conductivity divided by the charge. The correspondence extends even further: the rate of entropy production resulting from the thermodynamic treatment is the natural generalization of the I^2R relation, i.e. $eTs=JX=J^2L^{-1}=LX^2$. Also, the free path integrals of Sommerfeld-Bethe are consistent with $L=nb$, and Eq. (2).

to zero current, as was assumed, and to zero applied voltage.

The zero current condition is therefore interpretable as a kinetic equilibrium in which the flow produced by the gradient of the electrostatic potential is balanced by an opposite flow of equal magnitude due to the concentration gradient.[19] There is no conceptual difficulty here in giving separate meanings to the electrostatic potential and to the concentration-dependent term.

3. Boundary Conditions

The inner potential in a particular dielectric differs from that in the external vapor by the value of χ appropriate to the dielectric. This results in the correct Boltzmann conditions at the metal-dielectric boundary, and, since the difference between the Volta and inner potential is constant for one dielectric, the values of fields and charges may be determined by the usual classical relations. At a boundary, however, the thermodynamic conditions are satisfied. Thus, the requirement of constancy of Volta potential at the uncharged (and without double layer) contact between two dielectrics implies that there is a jump in the inner potential equal to the differences between the χ's of the two dielectrics. When this condition is applied to the expressions for the charge densities, there results the Boltzmann expression for the relation between charge densities in the dielectrics at the interface, the potential energy jump appearing in the numerator of the exponential. The usual expression for the continuity of the normal component of the displacement and the tangential component of the field is valid. The boundary conditions on the quantities in the dielectric remembering that φ and χ are positive downwards, are therefore:

(a) At a metal-dielectric interface,

$$\psi_i=\varphi_i-\chi_i; \quad n_i=N_e\exp[-(\varphi_i-\chi_i)/kT]. \quad (16.1)$$

(b) At a dielectric-dielectric interface,

$$\psi_i-\psi_j=-\chi_i+\chi_j; \quad \epsilon_iF_i-\epsilon_jF_j=4\pi\sigma. \quad (16.2)$$

(c) At a smooth maximum in one dielectric, the electric field becomes zero, and the inner potentials are equal, as the maximum is approached from either direction. (16.3)

These boundary conditions insure compliance with classical electrostatic theory and thermodynamic requirements at boundaries and interfaces. However, they do not yet specify the internal field in the dielectric, since this can have any one of an infinite number of values depending upon the value of g^2 and yet satisfy all other conditions. A specification of the value of the potential, electric field, or concentration at a particular value of x, or of the slope of any one of the quantities at a boundary or interface, determines g^2 and renders the whole set of values of all three quantities definite at all points within the dielectric.

[19] Earnshaw's theorem in classical electrostatics [J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill Book Company, Inc., New York, 1941), p. 116; G. P. Harnwell, *Principles of Electricity and Magnetism* (McGraw-Hill Book Company, Inc., New York, 1949), second edition, p. 63; or W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill Book Company, Inc., New York, 1939), first edition, p. 13] that a charged particle placed in an electrostatic field cannot be maintained in stable equilibrium under the influence of the electrostatic forces alone is not applicable. Both diffusion and electrostatic forces are acting, and the equilibrium is dynamic, being maintained by diffusion into and through the conduction levels of the insulator. The electrodes at either end of the insulator furnish the high concentration regions which maintain the external contribution to the dynamic equilibrium.

IV. THE FINITE AND INFINITE DIELECTRIC WITH SPACE CHARGE; ZERO CURRENT

1. The Finite Dielectric Between Two Metal Electrodes

A metal-adhesive-metal bond, or a plastic sheet between metal molding plates is represented by the model described in the foregoing title. For generality, it is assumed that the electrodes are of different metals. It is necessary to distinguish between the case in which the inner electrostatic potential has a maximum in the dielectric, and that in which it is monotonic within the dielectric, see Fig. 2. Borgnis (see reference 13) shows that these are the only two types of behavior possible. However, he assumes that the applied voltage is the mechanism responsible for the shift of the potential maximum to the boundary, and that, when the maximum reaches the boundary, the saturation current is attained. To see what the mechanism is on the thermodynamic picture, it is necessary to consider separately the expressions for a monotonic potential in the dielectric, and for a potential maximum in the dielectric. This is done in sections (a) and (b) hereafter. From Eqs. (17) and (20) developed there, and since (Table I, reference 2) a small energy gap at the boundary corresponds to a small x_0 or a large volume density of electrons at the boundary, it can be shown that:

1. If the energy gaps or x_0's at the two ends of the dielectric are equal, the circular function solution is valid for all thicknesses of dielectric. If the energy gaps are nearly the same, the circular function solution is valid for almost all thicknesses.
2. If the energy gaps or x_0's at the two ends of the

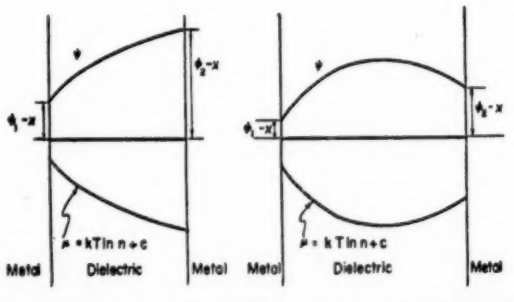


FIG. 2. Electrostatic potential (ψ) and chemical potential (μ) for the cases of monotonic potential and of potential maximum, within the dielectric.

dielectric are appreciably different, the circular function solutions are only valid if the dielectric thickness, D, is considerably greater than both x_0's. Otherwise, the infinite solution, or the hyperbolic solution must be used. Therefore, for large boundary electron volume densities, the circular function solutions will be appropriate if the dielectric is of reasonable thicknesses. Since, however, for energy gaps of, say, more than 1.5 volts, the x_0's are more than 10^5 cm, it is unlikely with metals of different work functions that the circular solutions will be appropriate for the usual dielectric thicknesses.

The value of g or a thus depends on the x_0's, and on the dielectric thickness; g or a determine the type of solution. Therefore, the total charge, charge distribution, field, or inner potential within a particular polymer or ceramic in contact with electrodes, and at a particular temperature, is unambiguously determined by the nature of the electrodes or metal surface, the depth of the conduction level in the dielectric, and the dielectric thickness.

(a) *Potential Maximum in the Dielectric*

If there is a maximum in the dielectric, it is reasonable to assume that the maximum is smooth; the required additional boundary condition on each of the solutions is then $F(\alpha)=0$ approached from either direction.

Only Eqs. (12) yield a smooth maximum in the dielectric. In general, two solutions are necessary to avoid inconsistency at the surface $x=D$. Let the maximum be at $x=\alpha$.

$$\left.\begin{aligned} &x\leqslant\alpha: \\ &\psi_1=2kT\ln\left[\frac{\operatorname{sing}_1(x+x_{01})}{\operatorname{sing}_1 x_{01}}\right]+\varphi_1-\chi \\ &F_1=-(2kT/e)g_1\operatorname{cotg}_1(x+x_{01}) \\ &n_1=(\epsilon kT/2\pi e^2)g_1^2\csc^2 g_1(x+x_{01}) \\ &x\geqslant\alpha: \\ &\psi_2=2kT\ln\left[\frac{\operatorname{sing}_2(D-x+x_{02})}{\operatorname{sing}_2 x_{02}}\right]+\varphi_2-\chi \\ &F_2=(2kT/e)g_2\operatorname{cotg}_2(D-x+x_{02}) \\ &n_2=(\epsilon kT/2\pi e^2)g_2^2\csc^2 g_2(D-x+x_{02}). \end{aligned}\right\} \quad (17)$$

The boundary condition on the potential has already been satisfied; the others are

$$F_1(\alpha)=F_2(\alpha)=0;\quad \psi_1(\alpha)=\psi_2(\alpha); \quad (17.1)$$

$$\begin{aligned} n_{01}&=(n_1)_{x=0}=N_c\exp[-(\varphi_1-\chi)/kT]; \\ n_{02}&=(n_2)_{x=D}=N_c\exp[-(\varphi_2-\chi)/kT]; \end{aligned} \quad (17.2)$$

which yield

$$g_1=g_2=g=\pi/(D+x_{01}+x_{02});\quad 2\alpha=D+x_{02}-x_{01}; \quad (17.3)$$

$$x_{0i}^2=(\epsilon kT/2\pi e^2 n_{0i})u_{0i}^2\csc^2 u_{0i};\quad \text{where}\quad u_{0i}=gx_{0i}. \quad (17.4)$$

As D becomes infinite g approaches zero, and the solution approaches that of Eq. (11). The maximum is inside the dielectric only if

$$|x_{02}-x_{01}|<D \quad (18)$$

so that if the energy gaps $(\varphi_i-\chi)$ at the two metal-dielectric boundaries are considerably different, the maximum moves out of the dielectric into the physically unreal region between it and the hypothetical infinite diffusing planes. If, therefore, the work functions of the metals are not very different a solution is possible in which the potential maximum is within the dielectric. If the work functions are equal, the maximum is at the center, and the two solutions can be represented by a single solution by moving the origin to the center as in the solution of Fowler.

(b) *Monotonic Potential in the Dielectric*

If the electrodes are of metals whose work functions are quite different, a single solution only is necessary. In this case, one electrode supplies all the charge in the dielectric and (except when α is at $x=0$ or D) additional charge to the other electrode by diffusion through the dielectric. Taking the first set of equations in Eq. (17) and their analogs from (11) and (12) as the solution, the boundary conditions are (17.2) and

$$\psi_1(D)=\varphi_2-\chi. \quad (19)$$

The equation which determines g is

$$[\operatorname{sing}(D+x_{01})]/\operatorname{sing}x_{01}=(n_{01}/n_{02})^{\frac{1}{2}} \quad (20)$$

and its analogs from (11) and (12). Thus, when the maximum is in the dielectric, the boundary conditions are applied on either side of the maximum, i.e., to the separate solutions; as the maximum moves out of the physically real region, the metal-dielectric contacts both occur on the same side of the maximum, and the boundary conditions apply to one solution only.

Nevertheless, it may be shown[20] that in this case, also, the value of g or of a should be $\pi/(D+x_{02}+x_{01})$; x_{02} is to be determined from (17.2) and (17.4) with $i=2$.

[20] In the case of the circular functions this is evident by applying the requirement that for each boundary the value of the expressions in the first half of (17) should be the same at that boundary whether the origin is taken initially at that boundary or, initially, at the other boundary and moved to the first boundary with the direction of the x-axis reversed. The same value must then be used in the argument for the hyperbolic functions to assure analytic continuation. This may also be shown directly since the circular and hyperbolic functions involved here are actually the first terms of an asymptotic expansion. The requirement in the first sentence of this footnote, taking account of Stokes' phenomenon for the asymptotic expansions of the Bessel functions involved, yields the value of a or g. This requires, of course, that x_{02} be given its value obtained from the last expression in the second half of Eq. (17) by letting $x=D$.

(c) *Charge Balance*

Assuming that metal and dielectric were uncharged before contact, the negative of the charge density on the surface of the metal will be equal to the total charge which has passed into the dielectric from that metal through unit area of surface. In the case of the potential maximum, charge has passed in from each metal surface, and the charge on either side of the inner potential maximum (i.e., of the charge density minimum) is that which has come from the electrode on that end. Therefore, if σ_M is the charge on the metal:

$$\left.\begin{aligned} -\sigma_{M0}&=\int_0^{\alpha} n dx=(\epsilon/4\pi)[F(\alpha)-F(0)] \\ -\sigma_{MD}&=\int_D^{\alpha} n dx=(\epsilon/4\pi)[F(\alpha)-F(D)] \end{aligned}\right\}, \quad (21.1)$$

where the second relation follows by Poisson's equation. Since α is the point of potential maximum, $F(\alpha)=0$. If, on the other hand, the dielectric is between metals of sufficiently different work functions, the monotonic case results, and the charge on the electrode of higher work function is that which has been furnished originally by the electrode of lower work function by diffusion through the dielectric. In this case:

$$\begin{aligned} -|\sigma_{M0}|+|\sigma_{MD}|&=\int_0^{D} n dx \\ &=(\epsilon/4\pi)[F(D)-F(0)]. \end{aligned} \quad (21.2)$$

Accordingly, the usual electrostatic expression for the surface charge on the metal and the field in the dielectric follows in both cases.

In the zero current case, if the appropriate Poisson, and Gauss' flux, relations are used, Eq. (9) becomes

$$g^2=\frac{2\pi e^2}{\epsilon kT}\left[N_0-\frac{2\pi\sigma_0^2}{\epsilon kT}\right], \quad (22)$$

in which σ_0 is the surface charge density on the metal in contact with the dielectric. By analogous considerations, the field at infinity in the solution (12) implies a surface charge density on the dielectric at infinity of amount $\sigma_\infty=\epsilon kTa/2\pi e$.

Surface (interface) states are unnecessary to the present explanation of the phenomena in dielectric-metal contacts. Suppose, however, that it appears likely that such states exist as, for example, by the transfer to the dielectric by tunnel effect of electrons of energy less than $\varphi-\chi$ above the Fermi level of the metal, and therefore unable to penetrate appreciably into the dielectric. In a subsequent paper, it will be shown that such intermediate energy levels do not affect the thermodynamic conditions on n at the boundary if the material is insulated throughout. If, however, it is desirable to introduce interface levels, the present treatment is still valid if the boundary is taken within the dielectric just short of the actual interface; conceptually the surface charge density on the metal should then be replaced by the algebraic sum of that charge density and the charge density in interface levels.

V. EQUILIBRIUM CHARGE DISTRIBUTION IN HIGH POLYMERS, CERAMICS, ETC.

1. Before Separation from the Metal

If the resulting charge density is fairly large, the energy gap ($\varphi-\chi$) must have been small at at least one metal contact. The approximation of the infinite solution may then be used; for the one side, Eq. (11) is used, and the course of the potential, and charge density is that shown on either side of the center in Fig. 4 of reference 2. The amount of charge is determined by the energy gap, the temperature and properties of the dielectric; the adhesion encountered upon removal from the metal mold will include an electrostatic component of the magnitude "N_0kT" in which N_0 is given by Eq. (1). A similar effect occurs on the other side. In reference 2 it is shown that adhesive forces of some magnitude may be encountered. If the energy gap, $\varphi-\chi$, is large, the finite solution must be used, since the distance x_0 becomes greater than the thickness of the dielectric. On the other hand, when this occurs, the charge density in the dielectric is small and in many cases can be neglected so that it is not necessary to consider the problem at all.

As an illustration, consider two cases, namely those in which $\varphi-\chi$ is 0.8 and 0.1 ev, respectively. With 0.8 ev, the charge density will decrease by 1 percent in about one mm of depth for average values of ϵ and T. Even at a centimeter into the material it is still over $\frac{1}{8}$th its value at the surface. However, a thick sheet of polymer about a yard square would contain only 2 esu of charge from earlier contact on the one side, or double that from both.[21] If the energy gap is 0.1 ev, the charge (while still a true volume charge) will be concentrated close to the surface, decreasing by 99 percent in between 10^{-2} and 10^{-3} cm. If electrical breakdown were not to occur, this would mean that the charge density within the volume near the surface would be some 178 esu per cm^2 of surface.

2. Edge Corrections

With the model assumed here, in which all quantities are independent of y and z, a block of insulating material with metal plates on each end is a closed system from which the electron gas cannot escape laterally. Therefore it is possible to consider high charge densities and inner potentials,[22] and apply the condition that the inner

[21] The charge per cm^2 of surface contained in the plastic from one metallic contact is $Q=e\int_0^{\infty} n dx=n_0 e x_0$. Values of n_0 and x_0 are given in Table I, reference 2.

[22] With even somewhat extreme values of the constants the inner potential will not be above that at the interface by more than 0.7–1.0 ev.

electrostatic potential and the Volta potential are separated by a fixed value, at all points. Similarly, the plane infinite lubricated slider bearing builds up a high pressure in the lubricant; Michell's treatment of the finite bearing shows that the pressure decreases because of flow over the lateral boundaries. In the molding of commercial wall tiles, or insulators, the surface area is so large in comparison with the thickness (or with the characteristic distance in cases in which there is appreciable charge content) that edge corrections are not important. In a long specimen of small cross-sectional area, edge corrections may be important, and the problem should be solved in two or three dimensions, with appropriate boundary conditions (see III, section 3) on the lateral area which is not in contact with the metal plates.

3. Separation of Polymer and Metal

Consider again the one-dimensional distribution. When the plastic is removed from the mold, the equilibrium between metal and plastic is disturbed. The charges in each create an external field in air, of the separating condenser type (different from that in the interior of the insulator), which increases rapidly with the distance of separation of the metal from the dielectric, and may cause corona breakdown of the air.[23] The ions produced as a result of the breakdown will drift to the charged plastic and at least in part neutralize the interior charges. Such neutralization is not complete, however, since the material is an insulator; a very long time will be necessary for the contained charges to diffuse outward to the surface and become susceptible to exact neutralization.

Contact between metal and insulating material is never complete[24] with the pressures ordinarily used. Consider again the case where $\varphi - x$ was small and a large quantity of electricity is in equilibrium in the plastic before separation, such as the example in the last paragraph. The portion not in contact with the metal will not acquire the charge, since the electron atmosphere outside a metal in air or vacuum is extremely diffuse (outside copper in air, the indicated density is less than 10^{-40} electron per cm^3), and there is little charge available to diffuse into the plastic. Such charge as penetrates into the regions not in contact diffuses laterally from the regions behind the areas of contact, hindered by the low conductivity of the insulator. Upon removal of the metal, the regions in which there is very high charge density acquire high electrical potentials near the surface; this produces electrical breakdown in the surface, and the characteristic patterns should be a group of paths penetrating from the high charge density regions into the uncharged spots. The location of the actual paths will depend upon the actual surface structure and composition and can be expected to give a sporadic appearance; however, the over-all character of the picture should be approximately that shown in the figures in reference 1.

The exact mathematical solution of the equilibrium situation after separation is indicated elsewhere, but since the densities in air are small, and since the solutions refer to equilibrium, for which the molding machine operator is unlikely to wait, they are not discussed here.

4. Energy Barrier between Metal and Insulator

The equations developed are quite general and could apply to ions as well as electrons or holes, provided ion diffusion could produce an appreciable space charge. This is unlikely in comparison to the known behavior of electronic charges.[25] On the other hand, it is evident that the conditions at a metal-dielectric boundary can be chemically quite complex, and that neither the metal nor the dielectric can be considered as inert materials merely providing a volume within which an electron gas is somehow created. The energy gap $\varphi - \chi$ can be possibly in some few cases, merely the difference between the true Fermi level in the metal and the conduction level in uncontaminated polymer. In such a case, the metal would furnish electrons to the polymer. This type of behavior is known; thus in the Russell effect,[26] a fresh metal surface placed against a photographic plate sensitizes the silver halide and, by transfer of electrons makes it possible to develop a picture of the portions in contact. Again (in the machining of metals), metal chips cut under water are found quantitatively to produce hydrogen peroxide.[27] Each unit area of surface has therefore furnished a definite number of electrons to the water; if the metal is exposed to the air before dropping it into the water, the transfer of electrons is smaller.

On the other hand, chemical reactions between the metal and dielectric atoms at or near the interface may alter the boundary energy level relationships. Thus, when zinc is placed in contact with a plastic containing chlorine in its structure, a dechlorination of the plastic in the vicinity of the metal is observed with the Cl going off in negative ionic form and leaving behind a double bond. This would leave a positively charged plastic and result in an energy level picture corresponding to a loss of electrons from the valence levels of the plastic. The mechanism discussed above should then be thought of in terms of the diffusion of holes into the insulator, and positive space charges would be found in it.

[23] This is discussed in a forthcoming article by the authors of reference 2.

[24] See, for example, Bobalek, LeBras, Powell, and von Fischer, Ind. Eng. Chem. **46**, 572 (1954), in particular, the last paragraph of the article.

[25] However, it is known that electron bombardment of a metal electrode de-gasses hydrogen ions from the interior.

[26] W. J. Russell, Proc. Roy. Soc. (London) **A61**, 424 (1897), and *seq.* A summary is given by G. L. Keenan, Chem. Rev. **3**, 95 (1920).

[27] L. Grunberg, Proc. Phys. Soc. (London) **B66**, 153 (1953). Zn, Al, Mg, Ni; in the case of Zn, the number of molecules of H_2O_2 formed was 1.5 times the number of Zn atoms exposed.

Some such mechanism (in addition to lack of metallic contact) is undoubtedly responsible for the alternating positively and negatively charged regions found by Woodland and Ziegler.[1] The metal cannot be regarded as a uniform surface of unvarying physical characteristics. This is well illustrated in the corrosion of ordinary metals; alternating portions of the surface area are electropositive and electronegative with respect to each other, and a couple is formed which favors the corrosion; localized corrosion regions or pits are formed. That such differences can be involved in plastic or ceramic contact is illustrated, for example, by the fact that vinyl chloride will etch metals thus removing electrons at the etched areas, and by the localized adhesion areas illustrated by the very precise techniques of reference 22. Although the exact localized chemical mechanism may be difficult to determine, nevertheless, that portion of the phenomenon responsible for the charging can be described in terms of the considerations above.

VI. REMOVAL OF THE CHARGE

Once the charges are in the plastic, their removal will require flow through the material. Therefore, there should be a correlation between the surface resistivity and the observed charges on manufactured articles, and with their dust-collecting properties or lack of ordered behavior in the textile loom. Woodland and Ziegler have shown that the dust-collecting tendencies are directly traceable to the static charge, and have covered the question of surface resistance in detail.

Surface neutralization by preferential ionic adsorption occurs in humid air or when ion sources are present; it is incomplete and removable by contact. Surface treatments such as sprays or washes will in general also be temporary and subject to the hazards of any mechanism which removes them at a later time. Antistaticizing treatments should therefore use materials capable of penetrating at least a short distance into the insulator, and neutralizing the charges directly, or rendering the penetrated volume sufficiently conductive to permit the volume charges to flow to the surface. In doing so, however, it is often important that the properties and appearance of the plastic not be harmed. It may be helpful to manufacture the plastic with a more conductive outer layer; this is not a complete solution since this layer can still be a source of charge diffusion into the insulating portion behind it.[28]

However, even if the surface is destaticized, later contact with other materials will again commence to put charges on it. Therefore, in some cases, such charging tendencies must be regarded as a natural accompaniment of otherwise desired properties, and other steps should be taken without relying on destaticizers for more than a limited time protection. Packaging, surface coating with conducting materials (which may later be grounded if necessary), provision of a suitably humid working atmosphere, and accessory sources of ions in the processing region may be useful at times. In the particular case of parts for electrical insulating use, the problem becomes very difficult, and the present treatment yields few helpful suggestions. By trial, it may be possible to find a mold liner such as a metal, a plastic, or a ceramic, for which $\varphi - \chi$ at the interface is large. This would reduce the contained charge in the commercial product. It would then be important to ensure that in the manufacturing process, there would be no subsequent contact of the part, before it has reached room temperature, with other metals, or, even at room temperature, with substances capable of triboelectric charging. Triboelectric charging is, in practice, a more complex process than that treated here.

[28] The effect of semiconducting layers between metal and dielectric will be treated in a separate paper.

Diffusion, Static Charges, and the Conduction of Electricity in Nonmetallic Solids by a Single Charge Carrier.*† II. Solution of the Rectifier Equations for Insulating Layers

SELBY M. SKINNER
Department of Chemistry and Chemical Engineering, Case Institute of Technology, Cleveland, Ohio

(Received June 15, 1954)

The discussion of the thermodynamic equilibrium of an electron atmosphere in a dielectric is extended to the current-carrying steady state. Under the assumption that the experimental behavior at zero current is continuous with that when there are finite electrical currents, it is found that forward and reverse currents show distinct behavior. Expressions are developed for the electrostatic potential, field, charge density, the current-applied voltage relation, and the capacity and rectifying behavior of dielectric layers. The non-ohmic conduction exhibits rectification, and departures from the Wagner relation are predicted, of the type observed experimentally, including turnover voltages. The contact charging of a dielectric is the zero-current limit of the non-ohmic behavior which causes rectification.

I. EFFECTS OF CURRENT FLOW

1. Current-Carrying State

WHEN the insulator is no longer in a condition of electrostatic charge equilibrium only, but is transmitting current:

(a) An applied voltage exists, so that U_2-U_1 must $=eV\neq 0$.

(b) The thermodynamical state of the system is not an equilibrium one, but a steady state one.

(c) The electron gas outside the second electrode is at a potential $\varphi_2-\varphi_1+eV$ with respect to that outside the first electrode. Appropriate changes must be made in the boundary conditions.

(d) The temperature will rise because of the j^2R heat unless provision is made for cooling. Since constant temperature has been assumed, it is necessary to assume that suitable cooling is provided. For small currents, it may be assumed that with natural cooling the observed behavior will be sufficiently well represented by assuming constant temperature.

(e) Since actually observed insulating layers (at least in small thicknesses) show rectifying properties, e.g., the copper oxide layer, the results should indicate such rectification.

2. Continuity of Physical Behavior

The equilibrium state discussed in I (with zero current flow) is not essentially different physically from one with small current flow. The solution of Eq. (9) for small current should reduce to the solution at zero current as the current flow becomes zero.[1] If this is possible, the validity of the whole range of solutions is intuitively more satisfying, since the two types of phenomena are thus directly related, and, for example, equilibrium charge densities at contact are the natural consequence and limiting result of the observed current flows at the interface between two materials when the net current becomes zero.[2]

The condition will therefore be applied that *there be no discontinuity in the potential, field, or charge density, as j is continuously changed from any finite value through zero to a finite magnitude of opposite sign.*

3. Solution of the Diffusion Equation for Nonzero Current[3]

Since the equation is nonlinear, considerably more complex interrelationships between the experimental parameters may be expected than in equations of the usual linear type. The experimental behavior of dielectrics varies sufficiently from sample to sample, and, for example, with thickness and temperature so that some such complexity is indicated.

Equation (9) of I has been obtained by an integration with respect to x, and therefore g^2 does not depend upon x, but it may well depend upon j, or T, or b. When there is a current, in analogy to the treatment in reference 3,

* Originally received, November 3, 1953. Received in final form June 15, 1954. Portions of this paper were presented at the same times and places as mentioned in reference 1 of the first paper of this series, J. Appl. Phys. **26**, 498, 1955, hereafter referred to as I.
† With the support of the Aeronautical Research Laboratory, Wright Air Development Center, U. S. Air Force.

[1] There exist in the literature statements that the equilibrium condition at zero current may differ by finite amounts from that approached as the current decreases to zero.
Such a metastable state is not considered here. With the equalizing tendency of diffusion and thermodynamic fluctuations, it would not be expected to persist.

[2] While the considerations are applicable to a part of the phenomena in triboelectric charging, they are not the complete explanation; it is not at all clear that equilibrium (and certainly not, constant temperature) exists in the instances usually observed. The subject is comprehensively covered by P. S. H. Henry, Science Progr. **41**, 617 (1951).

[3] A solution to Eq. (9) has been obtained by F. Borgnis, Z. Physik **100**, 478 (1936). However, his treatment is not suitable for the present purpose since he determined the values of his constants to agree with Ohm's law at infinity, considering only the electrostatic potential. An approximate treatment has also been given by Mott and Gurney (I, reference 5), and R. C. Prim, Phys. Rev. **90**, 753 (1953), Appendix II.

Eq. (9) of I is modified by the substitutions

$$\left.\begin{aligned} F&=\frac{2kT}{e}\beta y; \quad z=\beta(x+x_0)+\beta^{-2}a^2 \\ \text{or} \quad &=\beta(x+x_0)-\beta^{-2}g^2 \end{aligned}\right\}, \tag{1}$$

where

$$\beta^3=2\pi e^2 j/\epsilon bk^2T^2$$

becoming thereby a Ricatti equation

$$\frac{dy}{dz}\pm y^2+z=0. \tag{2}$$

The form of the solution depends upon whether z is positive or negative. If $z>0$,

$$y=-z^{\frac{1}{2}}\frac{AI_{-2/3}(w)+BI_{2/3}(w)}{AI_{\frac{1}{3}}(w)+BI_{-\frac{1}{3}}(w)}=-z^{\frac{1}{2}}R(w), \tag{3}$$

where $w=(2/3)z^{\frac{3}{2}}$ and the I_n are Bessel functions of imaginary argument. If $z<0$,

$$y=(\bar{z})^{\frac{1}{2}}\frac{AJ_{-2/3}(\bar{w})-BJ_{2/3}(\bar{w})}{AJ_{\frac{1}{3}}(\bar{w})+BJ_{-\frac{1}{3}}(\bar{w})}=(\bar{z})^{\frac{1}{2}}\bar{R}(\bar{w}), \tag{4}$$

where: $\bar{z}=-z$; $\bar{w}=(2/3)(\bar{z})^{\frac{3}{2}}$; the J_n are the Bessel functions of the first kind, of real argument. Only one constant in y is significant. While separate solutions are not additive, analogous relations may be found in the mathematical literature.[4]

Defining the charge density by Poisson's equation, the solutions for the electric potential energy, field, and charge density, ne, are for $z>0$,

$$\left.\begin{aligned} \psi&=2kT\ln[z^{\frac{1}{4}}\{AI_{\frac{1}{3}}(w)+BI_{-\frac{1}{3}}(w)\}]+\text{const} \\ F&=-(2kT/e)\beta z^{\frac{1}{2}}R(w) \\ n&=(\epsilon kT/2\pi e^2)\beta^2 z[R^2(w)-1] \end{aligned}\right\}, \tag{5.1}$$

and for $z<0$,

$$\left.\begin{aligned} \psi&=2kT\ln[\bar{z}^{\frac{1}{4}}\{AJ_{\frac{1}{3}}(\bar{w})+BJ_{-\frac{1}{3}}(\bar{w})\}]+\text{const} \\ F&=-(2kT/e)\bar{\beta}\bar{z}^{\frac{1}{2}}\bar{R}(\bar{w}) \\ n&=(\epsilon kT/2\pi e^2)\bar{\beta}^2\bar{z}[\bar{R}^2(\bar{w})+1] \end{aligned}\right\}. \tag{5.2}$$

By Eq. (1), the coefficient β, and therefore, the quantities z and w, depend upon the value of the current.

4. Evaluation of the Constants A and B

By the foregoing paragraph 2, Eqs. (5) must continuously approach the solutions given in Eqs. (10)–(12) of I, as j approaches zero. The evaluation of A and B to satisfy this condition requires a mathematical stratagem, but can be shown to be valid by considering the asymptotic or near-zero values, respectively, of the final expressions.

The result of the development in Appendix I is that the three types of solution for $g^2\gtrless 0$, corresponding to Eqs. (10)–(12) of I, are:

$$\left.\begin{aligned} \text{I.}\quad \psi-\psi_0&=2kT\ln\left[\frac{\bar{w}^{\frac{1}{3}}\bar{H}_2(\bar{w})}{\bar{w}_0^{\frac{1}{3}}\bar{H}_2(\bar{w}_0)}\right] \\ F&=-(2kT/e)\beta\bar{z}^{\frac{1}{2}}\bar{G} \\ n&=(\epsilon kT/2\pi e^2)\beta^2\bar{z}[\bar{G}^2+1] \end{aligned}\right\}, \tag{6}$$

where the subscript zero refers to the value at $x=0$, and

$$\bar{z}=-z=\beta^{-2}g^2-\beta(x+x_0); \quad \bar{w}=\tfrac{2}{3}\bar{z}^{\frac{3}{2}}=\bar{\xi}\left[1-\frac{2}{3}\frac{\bar{u}}{\bar{\xi}}\right]^{\frac{3}{2}};$$

$$\bar{\zeta}=-\zeta=\beta^{-2}g^2; \quad \bar{\xi}=\tfrac{2}{3}\bar{\zeta}^{\frac{3}{2}}; \quad \bar{u}=g(x+x_0); \quad \bar{G}=\bar{H}_1/\bar{H}_2;$$

$$\bar{H}_1=J_{\frac{1}{3}}(\bar{\xi})J_{2/3}(\bar{w})+J_{-\frac{1}{3}}(\bar{\xi})J_{-2/3}(\bar{w});$$

$$\bar{H}_2=J_{\frac{1}{3}}(\bar{\xi})J_{-\frac{1}{3}}(\bar{w})-J_{-\frac{1}{3}}(\bar{\xi})J_{\frac{1}{3}}(\bar{w}).$$

$$\left.\begin{aligned} \text{II.}\quad \psi-\psi_0&=2kT\ln\left[\frac{w^{\frac{1}{3}}H_2(w)}{w_0^{\frac{1}{3}}H_2(w_0)}\right] \\ F&=-(2kT/e)\beta z^{\frac{1}{2}}G \\ n&=(\epsilon kT/2\pi e^2)\beta^2 z[G^2-1] \end{aligned}\right\}, \tag{7}$$

where:

$$z=\beta^{-2}a^2+\beta(x+x_0); \quad w=\tfrac{2}{3}z^{\frac{3}{2}}=\xi\left[1+\frac{2}{3}\frac{u}{\xi}\right]^{\frac{3}{2}}; \quad a^2=-g^2;$$

$$\zeta=\beta^{-2}a^2; \quad \xi=\tfrac{2}{3}\zeta^{\frac{3}{2}}; \quad u=a(x+x_0); \quad G=H_1/H_2;$$

$$H_1=I_{-\frac{1}{3}}(\xi)I_{-2/3}(w)-I_{\frac{1}{3}}(\xi)I_{2/3}(w);$$

$$H_2=I_{-\frac{1}{3}}(\xi)I_{\frac{1}{3}}(w)-I_{\frac{1}{3}}(\xi)I_{-\frac{1}{3}}(w).$$

III. (a) For $z<0$: the expressions in Case I, just above, except that

$$g=0; \quad \bar{H}_1=J_{-2/3}(\bar{w}); \quad \bar{H}_2=J_{\frac{1}{3}}(\bar{w}); \quad \bar{w}=\tfrac{2}{3}[-\beta(x+x_0)]^{\frac{3}{2}}. \tag{8.1}$$

(b) For $z>0$: the expressions in Case II, just above, except that

$$a=0; \quad H_1=I_{-2/3}(w); \quad H_2=I_{\frac{1}{3}}(w); \quad w=\tfrac{2}{3}[\beta(x+x_0)]^{\frac{3}{2}}. \tag{8.2}$$

These solutions reduce, when $\beta\to 0$, to those of (I), Section II, 4, as follows:

Solution	Becomes solution in Section 4 of I, for
I	$g^2=-a^2>0$
II	$g^2=-a^2<0$
III	$g^2=-a^2=0$

and pass continuously into each other, if the constants in the potentials are properly chosen. Since the arguments of the Bessel functions must be real or imaginary

[4] For example, G. N. Watson, *Bessel Functions* (The Cambridge University Press, New York, 1948), revised edition, Chap. IV.

and not complex, in order to carry physical significance, the valid solutions are those in which z or $\bar{z}$ is positive. This imposes the conditions:

$$\text{Solution I:}\quad g^2/(x+x_0)>\beta^3>-\infty \qquad \text{III. a: } \beta^3<0$$
$$\text{II: } -a^2/(x+x_0)<\beta^3<\infty \qquad \text{b: } \beta^3>0.$$

The behavior of I in the negative current range or II in the positive current range is straightforward. From the point of view of an observer at $x=0$, the former is the forward current and the latter the reverse current as defined in rectification theory. At the limit of the other part of the range, the argument of the Bessel functions of either kind becomes zero and a continuous transition to the other type solution takes place.

The values in Eqs. (6)–(8) are independent of the choice of origin.[5] The difference in the expressions for positive and negative current is indicative of the physical difference between the forward and reverse currents in rectifying contacts. However, the solutions may not be accepted uncritically exactly as given above; considerations analogous to those in I must be used.

5. Electrochemical Potential

The fact that the electrical potentials are represented by logarithmic functions of functions of the space coordinates yields for the electrochemical potentials, $U=\psi+kT\ln(n)+\text{const}$:

$$U-U_0=kT\ln\Omega, \tag{9}$$

where Ω has the values

$$\frac{\bar{w}^{4/3}[\bar{H}_1^2(\bar{w})+\bar{H}_2^2(\bar{w})]}{\bar{w}_0^{4/3}[\bar{H}_1^2(\bar{w}_0)+\bar{H}_2^2(\bar{w}_0)]}, \tag{9.1}$$

$$\frac{w^{4/3}[H_1^2(w)-H_2^2(w)]}{w_0^{4/3}[H_1^2(w_0)-H_2^2(w_0)]}, \tag{9.2}$$

respectively, and corresponding terms in the cases in Eqs. (8.1) and (8.2).

The space rate of variation of U must have a definite value to produce a particular current, since $j=-nb\nabla U$.[6]

Repeating Eq. (14) of I, the externally applied potential is related to the values of U at the ends of the dielectric by:

$$eV=U_2-U_1. \tag{10.1}$$

The relationship between applied voltage and current may be obtained from Eq. (10.1) either by inserting the expressions for the U's or as follows:

$$\frac{dU}{dx}=-\frac{j}{nb};\quad U_2-U_1=-\int_0^D\frac{j}{nb}dx=eV. \tag{10.2}$$

Equation (10.2) shows that the applied voltage is at least of the first order in the current.

6. The Constant, g^2 or a^2

Using Poisson's equation, the relation (Eq. 9 in I) for g^2 becomes

$$\left(\frac{eF}{2kT}\right)^2-\left(\frac{2\pi e^2n}{\epsilon kT}\right)=\beta^2z=\frac{2\pi e^2j}{\epsilon bk^2T^2}(x+x_0)-g^2, \tag{11}$$

which is the generalization of Eq. (10.4) in I to the case of current flow. If both the field and charge density are known at a boundary for a particular current, g^2 is thereby determined; the x_0 which enters into Eq. (11) is evaluated by the condition on n at the boundaries, and all other quantities are measurable. If $(x+x_0)\neq 0$, current flow causes a change in the relation between field, charge density, and g^2. Using the solutions of Eqs. (6) to (8), and Eq. (11), it may be shown that $j=-nbdU/dx$, identically.

Experimentally, the charge density and potential at the boundary of the dielectric may be regarded as measurable quantities; the measurement of the electric field at the boundary, is, on the other hand, not a convenient procedure, and other means of determining the value of g^2 are desired.

7. Asymptotic Development

In Eqs. (6) to (9), the quantity ξ is proportional to j^{-1}. Since the substances considered are insulators, only small currents need be considered, and development in powers of the current requires an asymptotic development in terms of ξ^{-1}. Physically, it is obvious that the location of the inner potential maximum, and other

[5] A change of origin which decreases the value of x will increase the value of x_0 and leave unchanged their sum. This has been accomplished by the choice of the constant in Eq. (9) in I. Otherwise a change of origin would have required a different value of g^2 to return to a designation of the same experimental conditions. Independence of choice of origin is essential to physical reality, just as in the solution of a linear equation. Equation (9) likens the current flow, caused by diffusion and electric field, to that at a distance from a source plane of infinite charge density, whose distance behind the electrode is determined by the physical conditions of the problem, such as the energy gap at the interface; the description should be unaffected by whether the origin is taken at the electrode or away from it.

[6] This value may be distributed in any way between the space variation of the electrostatic potential and the space variation of the logarithm of the concentrations without changing the computed current. However, the considerations in reference 16 of I, and the assumption of the simultaneous validity of the diffusion equation and Poisson's equation provide the differentiation between how much of the gradient of the electrochemical potential is attributable to each of the two terms, and permit unambiguous expressions for the field and charge densities within any one material.

The question is not one of the reality of the Galvani potential; Gibbs, Guggenheim, and van Rysselberghe have shown that the attempt to define the Galvani potential as apart from the electrochemical potential is without meaning; exhaustive recent attempts to improve the situation, such as that in the third reference below do not make the Galvani potential operationally significant. The inner potential used here is that discussed in I, Sec. III, 1, defined with the use of the experimentally determinable χ.

J. W. Gibbs, *Collected Works* (Longmans Green and Company, New York, 1928), pp. 429–30; E. A. Guggenheim, *Thermodynamics* (North-Holland Publishing Company, Amsterdam, 1950), second edition. *International Committee on Electrochemical Thermodynamics and Kinetics: Proc. of the IIIrd Meeting, Berne, 1951* (Carlo Manfredi, Milan, 1952), especially Sec. IV, pp. 275–403.

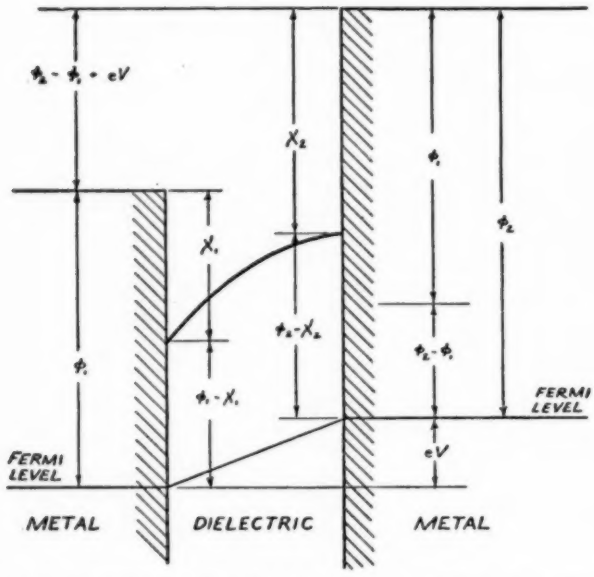


FIG. 1. Energy levels in a dielectric with metal electrodes, when voltage is applied.

parameters such as g, the x_0's, and the charge densities at the boundaries may depend upon the current. The development under these conditions is indicated in Appendix II. The results are:

$$\left.\begin{aligned} U-U_0&=kT\ln(A/A_0), \quad \psi-\psi_0=2kT\ln(B/B_0),\\ F&=-(2kT/e)C, \qquad\quad n=(\epsilon kT/2\pi e^2)D, \end{aligned}\right\} \tag{12}$$

where the quantities A, B, C, D, are power series in the current defined in Appendix II, and the subscript zero means that u is evaluated at the electrode-dielectric interface.

8. Finite Current-Carrying Dielectric

Since the solutions pass continuously into those for zero current described in I, it is to be expected that the energy level differences at the boundaries will affect the observed current-carrying behavior. It is again necessary to consider separately the case when the electrostatic potential is monotonic, and that when a potential maximum occurs in the dielectric.

(a) *Monotonic Potential*

Consider a dielectric specimen of thickness D, with, for example, identical energy gaps at each electrode. To be physically meaningful, it is necessary that the solution for small positive current at $x=0$ shows the same behavior as the solution for negative current of the same magnitude does at $x=D$, provided the origin is changed from $x=0$ to $x=D$, and the directions of the axis and the applied potential are reversed. If the electrodes are different, or for any other reason, the energy gap is different at the two electrodes, it is necessary to make appropriate adjustments of where the energy gaps are inserted into the boundary conditions when the origin and direction of the axis are changed.

If the contact is between the dielectric and a metal, it is unnecessary to consider changes in the Fermi level of the metal due to the current, except the raising of the level in one metal relative to the other by the amount of the applied voltage. This also causes the conduction level in the dielectric at that boundary to be raised by eV, so that the same boundary conditions on n are applicable as before; this is illustrated in Fig. 1. If an *equilibrium* existed, and the Fermi levels in both metals were at the same height, a conduction level at height $\varphi_2-\chi+eV$ would imply an equilibrium charge density in the dielectric of $N_c\exp[-(\varphi_2-\chi+eV)/kT]$; the applied voltage raises the Fermi level and increases the charge density to $N_c\exp[-(\varphi_2-\chi)/kT]$; the steady-state flow of current is the attempt to reach thermodynamic equilibrium by decreasing the charge density to the former value, unsuccessful because of the charges continually supplied by the source of external voltage.

The boundary conditions, therefore, will be taken as:

$$\psi_{x=0}=\varphi_1-\chi; \quad \psi_{x=D}=\varphi_2-\chi+eV;$$
$$n_{0i}=N_c\exp[-(\varphi_i-\chi)/kT];$$
$$g=g_0+\beta^3g_1+\cdots; \quad x_{0i}=x_{00i}+\beta^3x_{1i}+\cdots; \tag{13}$$

with the definitions of p and v_1:

$$\beta^3=2pj; \quad eV=2pjv_1kT;$$

in which only the first-order terms have been retained, and the electrostatic potential zero is at the Fermi level in the first metal. A single set of equations, Eqs. (12) suffice in the monotonic case. Applying these conditions: (1) the zero order terms are found to satisfy the zero-current expressions given in I, and, (2) the first-order terms yield

$$4g_0^4x_{1i}=4g_0^2(\tan u_{0i}-u_{0i})g_1 - 2u_{0i}\tan u_{0i}+\sin^2u_{0i}+u_{0i}^2. \tag{14}$$

The evaluation of g_1 requires additional considerations similar to those used in I in showing that g_0 is the same in the monotonic case as in the potential maximum case; it is not needed in the voltage-current relation however, since the same g_1 applies at both boundaries. The voltage-current relation is obtained by substituting the equations from Appendix II into Eq. (10.1), giving the analog of the Wagner equations, namely

$$\exp(eV/kT)-1=\sum_{i=1} h_ij^i, \tag{15}$$

$$h_1=pg_0^{-3}[\sin u_0\cos u_0-u_0-4g_0^2g_1]_1^2=2pv_1,$$

$$h_2=p^2(8g_0^6)^{-1}[16g_0^2g_1(u-\sin u\cos u-u\sin^2u) +3(1-2u^2+4u\sin u\cos u-2\sin^2u)+2u^2\sin^2u -16g_0^4x_{01}\sin^2u+2\sin^4u]_1^2.$$

The u's are evaluated at the boundaries as in Eq. (14), and the subscript and superscript on the brackets have the same meaning as in the expression of the value of a

definite integral. The coefficient h_1 may be written

$$h_1=\frac{e^2D}{g_0^2\epsilon bk^2T^2}[\{\sin(g_0D)/g_0D\}\cos(g_0s)-1], \quad (16)$$

where

$$g_0s=\pi(D+2x_{01})/(D+x_{01}+x_{02});$$
$$g_0D=\pi D/(D+x_{01}+x_{02}).$$

The quantity in brackets is always less than zero; by Sec. 4, this solution is valid if j is negative; therefore, Eqs. (15) and (16) correspond to the direction of easy flow. A similar treatment of Eqs. (II) yields

$$h_1=pg_0^{-2}D[\{\sinh(g_0D)/g_0D\}\cosh(g_0s)-1] \quad (17)$$

the solution being valid for positive j. This, therefore, represents the current-applied voltage relation in the reverse direction. The expressions (16) to (17) may also be obtained by inserting the zero-current approximation for n into Eq. (10.2), and higher terms in the series result from use in Eq. (10.2) of the higher terms in the expression for n in Appendix II. The higher terms give promise of explaining the turnover voltage observed in rectification by, for example, copper-oxide rectifiers. This will be discussed in Sec. 13 below.

The observed limit of resistivity of the material under steady-state conditions and uniform temperature as the current is decreased, is

$$\rho=V/jD=(e/kT)(\epsilon b)^{-1}[(D+x_{02}+x_{01})^2/\pi^2]\lambda, \quad (18)$$

where λ is the expression in brackets in Eqs. (16) or (17).

(b) *Potential Maximum*

As in I, it is necessary to use separate solutions on either side of the maximum; on the right-hand side (side 2), x is replaced by $D-x$, and the negative sign in front of the expression for the electrostatic field becomes a plus sign.

$$\psi_1=2kT\ln\left\{\frac{\sin u_1}{\sin u_{01}}[1+(pg_0^{-3}j/2)\{u_1-8g_0^2g_{11}+(4g_0^2g_{11}u_1-u_1^2+4v_1)\cot u_1\}_{x=0}^{x=x}]\right\}+\varphi_1-\chi$$

$$\psi_2=2kT\ln\left\{\frac{\sin u_2}{\sin u_{02}}[1+(pg_0^{-3}j/2)\{u_2-8g_0^2g_{12}+(4g_0^2g_{12}u_2-u_2^2+4v_2)\cot u_2\}_{x=0}^{x=x}]\right\}+\varphi_2-\chi+eV \quad (19)$$

in which, as in Appendix II,

$$u_1=g_0(x+x_{01}); \quad u_2=g_0(D-x+x_{02});$$
$$v_1=g_0^4(x_{11}+a_1); \quad v_2=g_0^4(x_{12}-a_1),$$

with analogous relations for the fields and charge densities. The boundary conditions are:

$$\left.\begin{array}{ll}\text{at } x=0: & (1)\ \psi_1=\varphi_1-\chi;\\ & (2)\ n_{01}=N_c\exp[-(\varphi_1-\chi)/kT];\\ \text{at } x=D: & (3)\ \psi_2=\varphi_2-\chi+eV;\\ & (4)\ n_{02}=N_c\exp[-(\varphi_2-\chi)/kT];\\ \text{at } x=\alpha: & (5)\ \psi_1=\psi_2;\ (6)\ F_1=0;\ (7)\ F_2=0;\end{array}\right\} \quad (20)$$

and in Eq. (19) the conditions on the potential at $x=0$, D, have been incorporated. The other five conditions applied to the zero order terms yield the same results as in the zero-current case treated in I. Applied to the first-order terms in the current (remembering that the term eV is of the first order in the current), they supply five relationships for the five quantities: x_{11}; x_{12}; g_{11}; g_{12}; and a_1. The term a_1 accounts for the shift of the potential maximum with increasing current, and since this shift is the same whether viewed from the right or from the left, it must enter into the v's in the manner shown in Eqs. (19). Applying the boundary conditions, and utilizing the results from the zero order terms, there is obtained a set of five simultaneous equations:

$$c_{i1}g_{12}+c_{i2}g_{11}+c_{i3}x_{12}+c_{i4}x_{11}+c_{i5}a_1=\delta_i,$$

in which

$$\begin{aligned}&c_{11}=-c_{12}=c_{21}=c_{32}=2\pi g_0^2; \quad c_{13}=c_{14}=c_{15}=c_{22}=c_{24}\\ &\quad=c_{31}=c_{33}=c_{42}=c_{44}=c_{45}=c_{51}=c_{53}=c_{55}=0;\\ &c_{23}=-c_{25}=c_{34}=c_{35}=4g_0^4;\\ &c_{41}=4g_0^2(1-u_{02}\cot u_{02}); \quad c_{52}=4g_0^2(1-u_{01}\cot u_{01});\\ &c_{43}=-4g_0^4\cot u_{02}; \quad c_{54}=-4g_0^4\cot u_{01}; \quad (21)\\ &\delta_1=\pi g_0^3v_1-(\pi/2)(\sin u_{02}\cos u_{02}\\ &\qquad-\sin u_{01}\cos u_{01}-u_{02}+u_{01});\\ &\delta_2=\delta_3=(\pi^2/4+1);\\ &\delta_4=u_{02}^2\cot u_{02}+\sin u_{02}\cos u_{02}+2u_{02};\\ &\delta_5=u_{01}^2\cot u_{01}+\sin u_{01}\cos u_{01}+2u_{01}.\end{aligned}$$

The first is obtained from boundary conditions 2, 4, and 5; the remainder, in order are obtained from the conditions 6, 7, 4, and 2 separately. The five simultaneous equations suffice to determine the five unknown parameters.

Since the electrical quanties are given by different expressions on the two sides of the maximum, the current-voltage relationship must be obtained as follows:

$$\begin{aligned}&\exp[(U_2-U_{\alpha 2})/kT]\\ &\qquad=1+[\sin u\cos u-u-4g_0^2g_1]_\alpha^2\\ &\exp[(U_1-U_{\alpha 1})/kT]\\ &\qquad=1+[\sin u\cos u-u-4g_0^2g_1]_1^\alpha \quad (22)\\ &eV/kT=\exp[(U_2-U_1)/kT]\\ &\quad=1+(1/3\xi_0)[\sin u_{02}\cos u_{02}\\ &\qquad+\sin u_{01}\cos u_{01}-(u_{02}+u_{01})+\pi]\end{aligned}$$

since $u_\alpha=\pi/2$; $u_{02}=g_0x_{02}$, $u_{01}=g_0x_{01}$. Therefore

$$\exp(eV/kT)-1=pg_0^{-2}jD[(\sin g_0D/g_0D)\cos g_0l+1], \quad (23)$$

where $g_0l=g_0(x_{02}-x_{01})$. Recalling from I that the location of the maximum is at $x=\alpha=[D+(x_{02}-x_{01})]/2$, it is possible to show that the current-voltage relationship Eq. (23) passes continuously into that given by Eq. (15) as the maximum passes either interface by variation of the x_0's. The choice between Eqs. (15) and (23) is

determined by whether the maximum is inside the dielectric at zero current.

To examine the behavior of the potential maximum as the current is increased from zero, it is necessary to examine a_1. Solving Eqs. (21), there is obtained

$$a_1=\Gamma/\Delta, \tag{24}$$

where

$$\Gamma=4t_1t_2(\pi+3s_1-s_2)+t_2[2\pi(3u_2-s_2)+4\pi(s_1+s_2)+4(1+4u_1^2+u_1s_1)+3\pi^2]-t_1[2\pi(3u_1-s_1)+4(1+4u_2^2+u_2s_2)+\pi^2]+[\pi^3+2\pi^2s_1+2\pi(s_2u_2-s_1u_1)+2(u_2^2-u_1^2)],$$

$$\Delta=16g_0^4(t_2+t_1+\pi),$$

$$t_i=\tan u_{0i}-u_{0i};\quad s_i=\sin u_{0i}\cos u_{0i}-u_{0i}.$$

If the electrodes are of the same material, $u_{02}=u_{01}$, and Eq. (24) becomes

$$a_1=[(8s+4\pi)t^2+8\pi st+4\pi^2t+2\pi^2s+\pi^3]/16g_0^4(\pi+2t). \tag{25}$$

In this case, a large energy gap at the boundary produces a very large x_0, compared to which the thickness of the dielectric is negligible. The quantities u_0 being $\pi x_0/(D+2x_0)$ may be represented by $(\pi/2)-\delta$ where $\delta\cong\pi D/4x_0$. The value of a_1 therefore, in the case of a large energy gap and little zero-current charging of the dielectric is

$$a_1=(2g_0^4)^{-1}[1+\pi^4D/64x_0+\cdots]. \tag{25.1}$$

If, on the other hand, the zero-current energy gap is small, so that the x_0's are small, and considerable charge transfer occurs at equilibrium, the value of a_1 is

$$a_1=(\pi^2/16g_0^4)[1-(2\pi^3/3)(x_0/D)^3]. \tag{25.2}$$

Thus, in both limiting cases, the location of the maximum is found approximately from

$$\alpha=\alpha_0+(2g_0^4)^{-1}\beta^3=(D/2)+g_0^{-4}pj,$$

so that its rate of transfer from the center of the dielectric, with current is

$$\pi e^2/\epsilon bk^2T^2g_0^4. \tag{25.3}$$

In the former case g_0 is approximately $\pi/2$, and in the latter case, it is approximately $\pi x_0/D$.

II. RECTIFICATION

9. Effect of the Boundary Energy Gaps on the Resistivity

The great variety of behavior of different dielectrics follows from Eqs. (18) and (23). The resistivity is inversely proportional to the charge carrier mobility, and depends both upon the energy gaps at the boundaries, and upon the thickness of the dielectric as related to the characteristic lengths x_0. Since, as in I, a small energy gap and large zero-current transfer of charge between electrode and dielectric implies a small x_0 which may range as low as 10^{-7} cm, and a large energy gap and a small charge transfer implies an x_0 of astronomical length, there would be expected considerable ranges of behavior. If both x_0's are small compared to D,[7] whether or not they are equal to each other, Eq. (23) gives for the resistivity,

$$\rho=\frac{e}{kT}\frac{1}{\epsilon b}\frac{D^2}{\pi^2}. \tag{26.1}$$

If both x_0's are large compared to the thickness of the dielectric, and both about equal, the resistivity is

$$\rho\cong\frac{e}{kT}\frac{2g_0^{-2}}{\epsilon b}=(4\pi n_0eb)^{-1} \tag{26.2}$$

the second equation being obtained by use of Eqs. (17) of I. In this case,

$$j\propto n_0ebF, \tag{26.3}$$

so that with a dielectric-metal combination in which large energy gaps exist between the electrode and the dielectric conduction levels, the observed behavior for small currents will approximate that in the case of a metal, but interpretation of the results as metallic conduction will give a slightly incorrect value of the product of charge density and mobility.

If both x_0's are large compared to the thickness of the dielectric, and quite different from each other,

$$\rho=\frac{e}{kT}\frac{1}{\epsilon b}\frac{D^2}{6}, \tag{26.4}$$

which is not essentially different from Eq. (26.1). Equation (26.4) also is valid if one x_0 is large compared to D and the other is small compared to D.

Intermediate values are more complex. The temperature dependence includes both a specific dependence upon T^{-1}, and the dependence of parameters such as the mobility and the quantities x_0 upon T, and requires special consideration for each case encountered experimentally.

10. The Applied Voltage-Current Characteristic

When the electrodes are identical, a zero-current potential maximum exists, and Eq. (23) is applicable. The Wagner relation becomes

$$\exp\left(\frac{eV}{kT}\right)-1=\frac{\pi e^2}{\epsilon bk^2T^2}\frac{D}{g_0^2}\left[\frac{\sin g_0D}{g_0D}+1\right]j, \tag{27}$$

showing that the voltage rises rapidly with current. The greatest slope occurs when g_0D is small, i.e., large energy gaps at the boundaries; the slope approaches zero, if $g_0D\to\pi$. Physically this means that, for a given thickness, a large initial charge transfer from the electrodes introduces a sufficient number of carriers so that the conductivity is large, whereas with a large energy gap at

[7] This is the limit of the solution I, and corresponds essentially to the solution IIIa.

the boundary, insufficient carriers exist to permit appreciable current without large voltages.

In the case of considerably different energy gaps at the boundaries, e.g., when the electrodes are of metals of quite different work functions, Eq. (17) is applicable. The slope of the analogous expression on the right-hand side is now negative, so that the voltage does not rise so rapidly with the current for the same absolute value of the coefficient of the current. However, the absolute value of this coefficient is now smallest when $g_0D\to 0$, and g_0l approaches zero. This occurs when x_{01} and D are small compared to x_{02}; the resistivity decreases not only with the thickness but also with a decrease of energy gap as before.

If an essentially infinite dielectric is considered, as in the second paragraph of the next section, the voltage current relation becomes in the direction of easy flow:

$$\exp(eV/kT)-1=c_1j-c_2j^2+\cdots; \qquad (28.1)$$

in the opposite direction:

$$\exp(eV/kT)-1=c_1j+c_3j^2, \qquad (28.2)$$

where $c_1=\frac{2}{3}(u_2^3-u_1^3)p$; $c_2=(p^2/9)(u_2^6+4u_1^3u_2^3-5u_1^6)$; $c_3=(p^2/9)(5u_2^6-4u_1^3u_2^3-u_1^6)$ and $u_2=D+x_0$, $u_1=x_0$. Considering only the first term on the right in either case,

$$|eV|=(2/3)\frac{\pi e^2}{\epsilon bkT}(D^3+3D^2x_0+3Dx_0^2)j. \qquad (29)$$

If the energy gap $\varphi_1-\chi$, is large, so that the dielectric is not appreciably charged at equilibrium, $x_0\gg D$, and remembering that for this case, $n_0=\epsilon kT/2\pi e^2n_0^2$ there is obtained $eV=jD/n_0b$ or its equivalent $j=n_0ebF$. If, on the other hand, the energy gaps are small, and appreciable contact charging occurs at equilibrium,

$$eV=(2/3)\frac{\pi e^2D^3j}{\epsilon bkT}=\frac{jD^3}{3bn_0x_0^2}.$$

It was shown in I that the total charge in the dielectric is n_0ex_0; letting this total charge equal CV where C is the capacity of the condenser formed by the electrodes and the dielectric, and substituting for x_0 in terms of n_0 by the equation first given,

$$V^2=\kappa j, \qquad (30)$$

where

$$\kappa=(D^3/3bC)\left[\frac{4\pi}{\epsilon kT}N_c\exp\{-(\varphi_1-\chi)/kT\}\right].$$

Since the capacity may be taken as constant, there should be observed a current proportional to the square of the applied voltage for small currents.[8]

11. Rectification Ratio

By the expressions developed, the slope on either side of zero current is the same, and the rectification ratio is one. This, however, does not mean that rectification cannot be obtained. In Sec. 4 it was shown that each type of solution, (I), and (II), is valid on both sides of zero current; on one side the solution is valid for any current, no matter how large, but on the other side the solution is valid only up to a particular current. At this current, the argument of the Bessel functions becomes zero, and further increase of current causes the argument to become pure imaginary, so that the analytic continuation of the solution involves Bessel functions of the other type. The slopes of the current-voltage curve are quite different on either side of this transition. When the transition occurs in the vicinity of zero current, the generally observed rectification behavior is predicted.

Consider, first, an insulator which is, for all practical purposes, infinitely thick. What is implied is not so much the actual thickness as the ratio of the thickness to the x_0's. In this case, g_0 approaches zero, and the solutions (8.1) and (8.2) are applicable. These solutions have not been treated in detail, since they are continuously approached as g_0 becomes zero, and accordingly all equations previously derived apply if g_0 is allowed suitably to approach zero. Both represent the same zero-current solution, corresponding to the infinite dielectric case in I and departures therefrom with increasing forward or reverse current are in opposite directions.

In the vicinity of zero current, they can be expanded in the usual infinite series for small argument. The voltage-current relations, Eqs. (28), imply slopes on either side of the equilibrium point of

$$c_1-j(c_1^2+2c_2);\ c_1-j(c_1^2-2c_3). \qquad (31)$$

The rectification ratio is therefore $1+2[(c_2+c_3)j]/c_1$ and increases with the current.

If, on the other hand, the thickness of the dielectric is not large compared to the x_0's the other solutions must be used. In this case, the transition occurs not at zero current but at $z=0$, or $\bar{z}=0$, for which the current has the value, at $x=0$, of

$$j=\frac{\epsilon bk^2T^2}{2e^2}\frac{\pi}{(D=x_{02}+x_{01})^2x_0}. \qquad (32)$$

Accordingly, in many cases, the transition occurs sufficiently near zero current so that for all practical purposes it is at zero current. The behavior in the vicinity of the transition will be treated in more detail later.

The hyperbolic functions (and Bessel functions of imaginary argument) do not ever, at zero or near zero current, represent the same solution as the circular functions (and Bessel functions of real argument). The same zero-current solution must be used for the current in both directions; small departures therefrom in either direction must be made with the asymptotic solutions previously discussed. In one direction the current may increase indefinitely without changing the type of Bessel functions. In the other direction, as the current

[8] Compare Mott and Gurney, reference 5 of I, p. 172.

TABLE I. Limiting values of the rectification ratio.

g_0D	g_0s	R	
0	0	1	1.0
	$\pi/2$	$\cosh(\pi/2)-1=$	1.5
	π	$[(\cosh\pi)-1]/2=$	5.3
	2π		∞
π	0	$[(\sinh\pi)/\pi]-1=$	2.7
	π	$[\sinh 2\pi/2\pi]-1=$	42
	2π	$[\sinh\pi\cosh 2\pi/\pi]-1=10^3$	

increases, there comes a point at which the transition must be made to Bessel functions of the other type, with, however, the same values of the zero order parameters. If different initial (zero current) conditions are encountered, the values of the parameters change, and a different solution must be followed through analytically.

An examination of the rectification ratio, $R=$ Eq. (17)/Eq. (16) shows that if the dielectric is thin, g_0D may approach zero, and then the maximum possible range of rectification ratios is encountered. The limiting cases are given in Table I.

The largest rectification ratio is obtained when x_{02} is much less than x_{01} and D is small. Therefore, the rectification efficiency is greatest for thin layers of dielectric, and considerable difference in energy gaps at the two boundaries. This energy gap may be controlled by the use of proper electrodes, or by the use of one electrode (or a layer upon the electrodes) which exhibits nonmetallic properties. Thus, the use of semiconducting electrodes or the treatments such as "forming" of the copper oxide and other rectifiers can be expected to produce necessary differences in energy gaps at the boundaries. In this connection, reference should be made to the discussion on the nature of the energy gaps in actual contacts which was given in I, Sec. V, 4.

It is obviously impossible to obtain zero thickness of useful dielectric, and if the dielectric thickness is reduced below about 10^{-5} or 10^{-6} cm, the tunnel effect enters so that complete diffusion equilibrium cannot be attained, and the considerations used here commence not to apply. However, such thicknesses are not desirable, even apart from considerations of chemical behavior and homogeneity, for it is known that the tunnel effect produces the opposite behavior from the diffusion equilibrium and therefore rectifying properties would be decreased.

A small layer of dielectric between a point contact and a semiconductor would increase the measured rectification ratio; this could help explain some of the anomalies in slope and shape of curve observed in such cases.[9]

[9] H. K. Henisch, *Metal Rectifiers* (Clarendon Press, Oxford, England, 1941); H. C. Torrey and C. A. Witmer, *Crystal Rectifiers* (McGraw-Hill Book Company, Inc., New York, 1948); Blakemore, DeBarr, and Gunn, *Semiconductor Circuit Elements, Reports on Progress in Physics* (Physical Society, London, 1953), Vol. XVI; M. F. Manning and M. E. Bell, Revs. Modern Phys. **12**, 215 (1940); and many others.

As the energy gaps at the boundary become more nearly equal, the rectification decreases, so that the initially prepared rectifier would not be expected to show efficient performance. This discussion has dealt with the ratio only; if the slope is large in the easy flow direction, a large voltage is required for a given current, and rectifier heating and other effects may make the dielectric unsuitable.

12. Capacity of the Rectifying Junction

The capacity of the dielectric specimen per unit area may be defined as

$$C=\left|\frac{Qe}{eV}\right| \text{ and since } Q=\int_{x=0}^{D} n\,dx,$$
$$C=\frac{\epsilon}{4\pi}\left|\frac{F_2-F_1}{eV}\right|_{x=0,\,D} \text{ or } \frac{\epsilon}{4\pi}\left|\frac{F_2+F_1}{eV}\right|_{x=0,\,D}, \tag{33}$$

which for small current, in the easy flow direction, becomes

$$C=\frac{\epsilon g_0}{4\pi}\left[\frac{\sin(u_{02}\pm u_{01})}{\sin u_{02}\sin u_{01}}\right], \tag{34}$$

in which the positive sign is taken if there is a potential maximum. The extension to the first term in j is obvious from Appendix II.

13. Turnover Voltage (Fig. 2)

The current enters into the voltage-current equation in an infinite series. Considering only the first two terms, we have

$$eV/kT=|h_1|\,j\pm|h_2-(1/2)h_1^2|\,j^2,$$

so that a voltage maximum in the current occurs if $(h_2/h_1)-(h_1/2)<0$, and $j=h_1/(2h_2-h_1^2)|$. This explains the turnover voltage usually observed, Fig. 2. To determine the location of the maximum, it is neces-

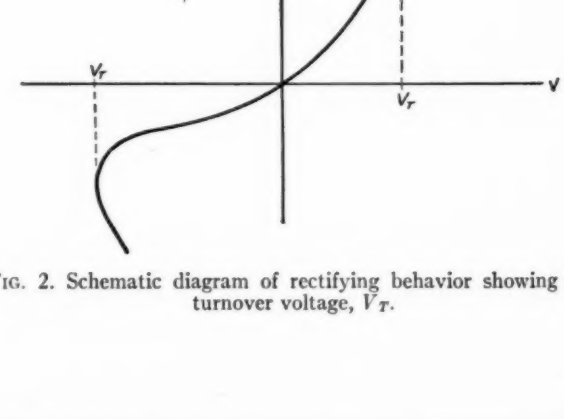


FIG. 2. Schematic diagram of rectifying behavior showing the turnover voltage, V_T.

sary to insert the expressions for h_1 and h_2, using the correct values for g_1 and a_1, as the particular case requires. This is a lengthy procedure, and since the succeeding terms in the series will modify the location of the potential maximum, it will not be undertaken here. The qualitative prediction of a voltage turnover, is, however, evident, and may be illustrated for the infinite dielectric, by Eqs. (31).

In this case only one turnover voltage is found; this occurs for

$$j=(3\epsilon bk^2T^2/2\pi e^2)\Omega,$$

where for small x_0, $\Omega\cong D^{-3}$, and for large x_0/D, $\Omega\cong(6x_0^3)^{-1}$. In the former case the turnover voltage is approximately $kT/e=39$ volts; in the latter, it is at $(D/2x_0)(kT/e)\cong 0$ volts.

14. Discussion

The electric field, for a given current is nonlinear and has a steep gradient in the vicinity of the electrode with the small energy gap, the steepness being independent of the current at small currents. Most of the potential drop measured in the dielectric between zero-current probes applied to the rectifying junction will be in the immediate vicinity of the interface.

The relationships pictured by the equations are those due to the behavior of space charges in thermodynamic equilibrium both throughout the dielectric, and with other space charges in the electrodes. The rectification is, of course, due to the piling up of the space charges towards one or the other electrode; the special behavior in the case of initially equal energy gaps is related to the symmetry of the initial electrical conditions. The energy gaps at the boundaries and the applied voltage enable the distinction between forward and reverse current behavior. Since only one carrier has been assumed, the equations are not applicable at large applied voltages when intrinsic charges appear. With the usual insulators, however, the *intrinsic* energy gap is so large that this is not likely to occur with applied voltage gradients short of internal breakdown. It has not been necessary to assume image forces or any other detailed characteristic of the interface, except the presence at either end of the dielectric of materials of suitable energy gaps. The energy gaps may be due to the contact potentials of the original metals and the conduction level of the original dielectric or may depend upon energetic considerations and chemical reactions between the metal and dielectric at the interface, but may be regarded as unambiguously specifiable.

Certain conclusions follow: (1) The *intrinsic* gap should be large enough so that smoothing effects from the thermal release of intrinsic electrons do not wash out the rectification. (2) The best rectification is obtained when the initial energy gaps at the boundaries are quite different. (3) There is a particular thickness of dielectric which will give the best rectification, for each combination of dielectric and electrodes. (4) In general, the best rectification will be obtained with quite thin, but not too thin layers of dielectric. (5) Since the resistance of the dielectric becomes great if its conduction level is much higher than both initial Fermi levels of the metal electrodes, it is desirable that the energy gap be not too large at one boundary. (6) Observed contact charge distributions in insulators are the limiting case of the current-carrying behavior which exhibits rectification, and both are due to the non-ohmic thermodynamic behavior of the charges in the insulator.

APPENDIX I. EVALUATION OF THE CONSTANTS IN THE NONZERO CURRENT SOLUTION

Because of Eqs. (5) it is immaterial whether the evaluation of A and B is accomplished by ψ, F or n, and on account of its greater simplicity, the expression for ψ will be used. Considering first, Eq. (5.1), and defining $u=a(x+x_0)$, as in I, there results by Eq. (1)

$$z=(\beta u/a)+\beta^{-2}a^2;\quad w=\tfrac{2}{3}z^{\frac{3}{2}},\tag{A1}$$

and A and B are to be evaluated so that Eq. (5.1) reduces to Eq. (12.1) of I as β approaches zero. Disregarding the additive constants for the time being, this requires that

$$\left.\begin{aligned}&\lim_{\beta\to 0}\varphi_1=\sinh u\\ \text{where}\quad&\\ &\varphi_1(u)=w^{\frac{1}{2}}[AI_{\frac{1}{3}}(w)+BI_{-\frac{1}{3}}(w)]\end{aligned}\right\}.\tag{A2}$$

An expansion in terms of β is not useful, but if φ_1 is expanded as a Taylor's series in u, there is obtained

$$\varphi_1=\sum_n \varphi_1^{(n)}(0)\cdot\frac{u^n}{n!};\quad\left(\varphi_1^{(n)}=\frac{d^n\varphi_1}{du^n}\right),\tag{A3}$$

where

$$\varphi_1(0)=\xi^{\frac{1}{2}}[AI_{\frac{1}{3}}(\xi)+BI_{-\frac{1}{3}}(\xi)],\tag{A3.1}$$

$$\xi=\tfrac{2}{3}\beta^{-3}a^3=\tfrac{2}{3}\zeta^{\frac{3}{2}},\tag{A3.2}$$

$$\varphi_1'(0)=\xi^{\frac{1}{2}}[AI_{-2/3}(\xi)+BI_{2/3}(\xi)],\tag{A3.3}$$

$$\varphi_1''(0)=(3/2)^{\frac{1}{3}}(\beta/a)^2\xi^{\frac{1}{2}}\varphi_1(0),$$

$$\varphi_1'''(0)=(\beta/a)^3\varphi_1(0)+(3/2)^{\frac{1}{3}}(\beta/a)^2\xi^{\frac{1}{2}}\varphi_1'(0),$$

and, in general, if

$$\varphi_1^{(n-1)}=X_{n-1}\cdot\varphi_1+Y_{n-1}\cdot\varphi_1',$$

then

$$\varphi_1^{(n)}=\left[(\beta/a)\left(\frac{3w}{2}\right)^{\frac{1}{3}}\frac{dx_{n-1}}{dw}+(\beta/a)^2\left(\frac{3w}{2}\right)^{\frac{2}{3}}Y_{n-1}\right]\varphi_1$$
$$+\left[(\beta/a)\left(\frac{3w}{2}\right)^{\frac{1}{3}}\frac{dY_{n-1}}{dw}+X_{n-1}\right]\varphi_1'.$$

From the foregoing expressions, there may be obtained by induction, and by the series expansion of $\sinh u$, the

necessary conditions desired. That they are sufficient conditions also, may be shown by asymptotic expansion of the resulting expression for the electrostatic potential, namely Eq. (7). These conditions are

$$\varphi_1(0)=0, \quad \varphi_1'(0)=1, \tag{A4.1}$$

and

$$\lim_{j\to 0}(\beta/a)^2\zeta=1, \quad \lim_{j\to 0}(\beta/a)^3=0. \tag{A4.2}$$

By the definition of ζ in Eq. (A3.2) $(\beta/a)^2\zeta=1$, and by the definition of β, $\lim_{j\to 0}[\beta/a]=0$ if $a\neq 0$. Equations (A4.2) are therefore satisfied. Using the well-known Wronskian relation

$$I_{-n}(w)I_{n-1}(w)-I_n(w)I_{-n+1}(w)=\frac{2\sin n\pi}{\pi w},$$

there are obtained, after solving Eqs. (A4.1) for A and B

$$\left.\begin{aligned} A&=c\xi^{\frac{1}{2}}I_{-\frac{1}{3}}(\xi)\\ B&=-c\xi^{\frac{1}{2}}I_{\frac{1}{3}}(\xi)\end{aligned}\right\}, \tag{A5}$$

where the constant is the same in both expressions and may be absorbed in the additive constant in Eq. (7). An analogous procedure and Wronskian yields, in the case of the Bessel functions of real argument, i.e., Eq. (5.1)

$$\left.\begin{aligned} A&=c_1\bar{\xi}^{\frac{1}{2}}J_{-\frac{1}{3}}(\bar{\xi})\\ B&=-c_1\bar{\xi}^{\frac{1}{2}}J_{\frac{1}{3}}(\bar{\xi})\end{aligned}\right\}. \tag{A6}$$

In the intermediate case, when $a^2=0$, continuity with the cases $a^2\neq 0$ is obtained by the choice $B=0$. When these values of the constants are used, the expressions for all quantities, electrostatic and electrochemical potential and potential differences, field and charge density, are continuous through $j=0$ and $a^2=g^2=0$.

APPENDIX II. ASYMPTOTIC DEVELOPMENT OF THE EXPRESSIONS FOR THE ELECTRICAL QUANTITIES

Since the solutions of Eq. (7) may be obtained from those of Eq. (6), by the substitutions $g=-ia$, and $\xi=i\bar{\xi}$, only the former will be developed. With obvious adjustments, these become the expansions about $j=0$ of Eqs. (8). The fact that the latter is a convergent series about $j=0$, suggests that Eqs. (12) and its analog may be a convergent series. This is not taken up here, but it may be noted that the expressions in (B2) below and its analog may be differentiated with respect to x, and the corresponding next lower expression is obtained, after multiplication by the appropriate constants which follow from the electrostatic laws.

All Bessel functions are defined as in reference 10. Using asymptotic expansions of the Bessel functions $J_{\frac{1}{3}}$ etc., of the Stokes type, there results

$$\left.\begin{aligned} H_1&=P\cos(w-\xi)-Q\sin(w-\xi);\\ H_2&=R\cos(w-\xi)-S\sin(w-\xi),\end{aligned}\right\} \tag{B1}$$

where

$$\pi(w\xi)^{\frac{1}{2}}P=3^{\frac{1}{2}}\left[1+\frac{5\cdot7\cdot13}{2(72w)^2}-\frac{5\cdot7\cdot11}{2(72\xi)^2}-\frac{5\cdot7}{(72w)(72\xi)}+\cdots\right],$$

$$\pi(w\xi)^{\frac{1}{2}}Q=3^{\frac{1}{2}}[(7/72w)+(5/72\xi)+\cdots],$$

$$\pi(w\xi)^{\frac{1}{2}}R=3^{\frac{1}{2}}[(5/72w)-(5/72\xi)+\cdots],$$

$$\pi(w\xi)^{\frac{1}{2}}S=3^{\frac{1}{2}}\left[1-\frac{5\cdot7\cdot11}{2(72w)^2}-\frac{5\cdot7\cdot11}{(72\xi)^2}+\frac{5\cdot5}{(72w)(72\xi)}+\cdots\right].$$

Next there are obtained appropriate expressions for w and ξ; it is necessary to remember that since the location of the potential maximum may depend upon the current, and boundary conditions will be encountered there, x must also be treated as a function of the current in the general development. Therefore, let

$$g=g_0+\beta^3g_1+\beta^6g_2+\cdots; \quad x=\bar{x}+\beta^3a_1+\cdots;$$
$$x_{0i}={}_0x_{0i}+\beta^3x_1+\cdots; \quad u_{0i}=g_0({}_0x_{0i}+\bar{x});$$
$$v=g_0^4(x_1+a_1); \quad \xi_0=\tfrac{2}{3}\beta^{-3}g_0^3.$$
$$(3\xi_0)^{-1}=pg_0^{-3}j \quad \text{where} \quad p=\pi e^2/\epsilon bk^2T^2.$$

Then

$$w=\xi_0[1+(1/\xi_0)(2g_0^2g_1-u_0)+\cdots];$$
$$\xi=\xi_0[1+(1/\xi_0)(2g_0^2g_1)+\cdots];$$
$$w-\xi=-u_0+(1/6\xi_0)(-4g_0^2g_1u_0+u_0^2-4v),$$

and after some manipulation, there is obtained for the quantities in Eq. (12):

$$\left.\begin{aligned} A&=1+pg_0^{-3}[\sin u_0\cos u_0 - u_0-4g_0^2g_1]j+\cdots,\\ B&=\sin u_0+(pg_0^{-3}/2)[(u_0-8g_0^2g_1)\sin u_0 + (4g_0^2g_1u_0-u_0^2+4v)\cos u_0]j+\cdots,\\ g_0^{-1}C&=\cot u_0+(pg_0^{-3}/2)[(4g_0^2g_1-2u_0)\cot u_0 + (u_0^2-4v-4g_0^2g_1u_0)\csc^2u_0+1]j+\cdots,\\ g_0^{-2}D&=\csc^2u_0[1+(pg_0^{-3})\{(4g_0^2g_1-2u_0) + (\sin^2u_0+u_0^2-4g_0^2g_1u_0-4v) \times\cot u_0\}j+\cdots].\end{aligned}\right\} \tag{B2}$$

At the metal dielectric interfaces both $\bar{x}$ or $D-\bar{x}$, and a_1 are taken as zero.

[10] H. Jeffreys and B. S. Jeffreys, *Methods of Mathematical Physics* (Cambridge University Press, Cambridge, England, 1950).

Small Particle X-Ray Scattering by Fibers, Size and Shape of Microcrystallites

A. N. J. HEYN*
Clemson College, Clemson, South Carolina
(Received May 14, 1954)

A theoretical and experimental analysis is given of small particle x-ray scattering by fibers.

The scattering by different fibers of high orientation is compared with the theoretical scattering curves derived for independent scattering by an assemblage of parallel cylinders. Under certain conditions of swelling, excellent agreement is obtained between experimental and theoretical curves. From this agreement it is concluded that under these conditions, independent particle scattering is realized and that interparticle interference and multiple scattering do not play a part.

Under all other conditions (dry fiber and other conditions of swelling) interparticle interference plays an important role in the scattering by these "densely packed" materials as can be concluded from the different distributions of intensity. The swelling method offers a means for separating independent and nonindependent scattering.

An analysis of the influence of various conditions of swelling is made, and the correct experimental conditions for obtaining independent scattering are derived from the results.

From the scattering under such conditions, the particle size in different fibers is determined on the basis of the theoretical formulas for independent scattering.

INTRODUCTION

X-RAY scattering at small angles by fibers has been explained in previous communications[1] as "small particle scattering" and has been related to the microcrystalline structure of the fiber. Information about shape, size, distribution, and orientation of the microcrystallites in the fiber may, in principle, be obtained, therefore, from the distribution of scattered intensity.

One of the essential features which govern the scattering by these materials and which offers the main difficulty in its analysis and application was found to be the close packing of the particles. The microcrystallites of the fibers are so close together that the radiation from neighboring particles generally interferes. As a result, the independent particle scattering or "form factor scattering" will be obscured and complicated. This complication makes it very difficult to analyze the scattering and evaluate it in terms of particle size, for which independent scattering is wanted.

This circumstance is also the main problem in the interpretation of the small particle scattering by many other materials. Some authors[2–4] assume that "form factor scattering" is essential, and that interference between particles can be neglected; others[5,6] believe that interference between neighboring particles or even diffraction by "Bragg reflection planes"[5] is essentially responsible for the scattering. These extreme views have been brought to a point in a controversy between Kratky and Hosemann.[7] Lund and Vineyard[8] emphasize on theoretical grounds the difficulties of determining particle sizes when appreciable interparticle interference is present. Their point of view will be confirmed in the following for the case of fibers, but at the same time a way out of these difficulties will be shown, at least with regard to fibers.

The present communication brings an analysis of the theoretical and experimental aspects of x-ray scattering by *fibers* with special reference to the separation of independent and nonindependent scattering and the conditions under which such a separation can be experimentally achieved. The experimental isolation of "form factor scattering," from the effect of interparticle interference, furnishes a way to overcome the foregoing difficulties, and to gain information from the scattered intensity about the nature and size of the particles on the one hand and their distribution on the other hand.

The analysis is carried out by a direct comparison between the scattering calculated for certain theoretical models and the scattering actually observed under various conditions in experiment.

THEORY OF SMALL PARTICLE SCATTERING BY FIBERS

The fibers studied in this paper have a high degree of orientation, the microcrystallites being parallel to the

* Professor of Natural and Synthetic Fibers, School of Textiles.

[1] A. N. J. Heyn, J. Am. Chem. Soc. **70**, 3138 (1948); Textile Research J. **19**, 163–72 (1949); J. Am. Chem. Soc. **72**, 2284–5 (1950a); J. Am. Chem. Soc. **72**, 5768–9 (1950b); Nature **172**, 1000 (1953); Textile Research J. **23**, 782–87 (1953).

[2] A. Guinier, Proc. Phys. Soc. (London) **57**, 310–24 (1945); J. chim. phys. **40**, 133–150 (1943); Ann. phys. **12**, 161–237 (1939); *Radiocrystallographie* (Dunod, Paris, 1942) Chapter 7.

[3] R. Hosemann, Z. Physik **113**, 751–68 (1939); Z. Physik **114**, 133–69 (1939); Z. Elektrochem. **46**, 535–55 (1940); Kolloid-Z. **117**, 13 (1949); Ergeb. exakt. Naturwiss. **24**, 143–221 (1951).

[4] C. G. Shull and L. C. Roess, J. Appl. Phys. **18**, 295–307 (1947); J. Appl. Phys. **18**, 308–13 (1947).

[5] O. Kratky, Naturwiss **26**, 94 (1938); Z. Elektrochem. **46**, 556 (1940); Naturwiss. **30**, 542–43 (1942); Kolloid-Z. **120**, 24–39 (1951); O. Kratky and F. Schoszberger, Z. physik. Chem. **B39**, 145–54 (1938); Kratky, Sekora, and Treer, Z. Elektrochem. **48**, 587–601 (1942); O. Kratky and A. Wurster, Z. Elektrochem. **50**, 249 (1944); O. Kratky and G. Porod, J. Colloid Sci. **4**, 35–70 (1949); Kratky, Kriegl, and Porod, Z. Elektrochem. **50**, 146–64 (1952).

[6] G. Porod, Kolloid-Z. **124**, 83 (1951); Kolloid-Z. **125**, 51 (1952).

[7] O. Kratky, Z. Elektrochem. **46**, 556 (1940).

[8] L. H. Lund and G. H. Vineyard, J. Appl. Phys. **20**, 593–7, (1949); G. H. Vineyard, J. Appl. Phys. **74**, 1076–83 (1948).

fiber axis (or nearly so). The scattering obtained consists of a narrow band along the equator of about the same width as the height of the beam (Fig. 3). The crystallites, therefore, may be concluded to be very long, as compared to their radius and the wavelength of the radiation. For the calculation, these particles may be considered to have infinite length. The scattering theory to follow is based, therefore, on a model consisting of a system of parallel cylindrical rods which scatter either independently or nonindependently, and which have nearly constant diameter and homogeneous inner-structure.

The amplitude ϕ scattered by such a particle is proportional to

$$\frac{2J_1(kR)}{kR}, \tag{1}$$

in which $k=4\pi \sin\theta/\lambda$, θ being half the scattering angle, and R the radius of the cylinder. For small-angles $4\pi \sin\theta/\lambda = 2\pi\epsilon/\lambda$, ϵ being the scattering angle in radians. [The polarization term $\frac{1}{2}(1+\cos^2\theta)$ is not included since the scattering takes place at small-angles where this term is nearly unity.] The function $\phi(kR)$ may be regarded as the form factor of the particle in analogy with the atomic form factor. The scattered intensity is proportional to the square of this function.

If instead of one single cylinder, an assemblage of parallel cylinders is present, and the elementary waves from the single particles do not interfere (*independent or single particle scattering*), and if the cylinders have uniform size,† the total radiation is the sum of the radiations from the separate cylinders, or $N\phi^2(kR)$ in which N is the number of cylinders.

The theoretical scattering curve for an assembly of independent scattering particles will, therefore, be represented by the function $[2J_1(x)/x]^2$ given in Fig. 1. In this function, x is to be substituted by kR. This curve reaches half the intensity value at $kR=1.61$, a first zero value at $kR=3.83$, a second maximum, with an intensity of 0.0175 of the zero angle intensity, at $kR=5.10$, a second zero value at kR 7.00, etc. The intensities of further maxima decrease more and more and, therefore, only the first maximum is given in Fig. 1. The theoretical curves for cylinders of a few different average radii are given in Fig. 2.

If the radiation of the neighboring cylinders does interfere (*nonindependent scattering, or interparticle interference*), the total radiation is obtained by adding the radiations from the individual cylinders together, but this time taking into account the phase differences.

The scattering function may be obtained in analogy with the procedure for the scattering by molecules in a

† If the diameter of the cylinders is subject to considerable variation, a factor for mass distribution has to be introduced into the formula. This can be done following the general procedures of Hosemann (see reference 3) and Shull and Roess (see reference 4). This influence will be much smaller in the case of cylindrical particles than in the case of spherical particles and is neglected in the present publication.

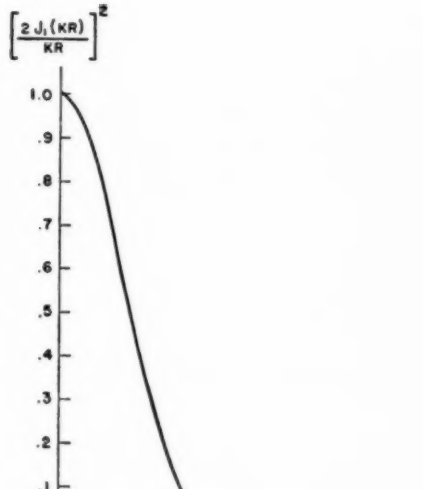


FIG. 1. Theoretical scattering curve for an assemblage of parallel cylinders of uniform size. Ordinate: absolute intensity. Abscissa: $kR=2\pi\epsilon/\lambda$.

fluid (Zernike and Prins,[9] Debye and Menke,[10] and Gingrich[11]). The same procedure has been followed by Lund and Vineyard[8] and Oster and Riley[12] in theoretical discussions of scattering by small particles.

In the case of fibers, the molecules of the fluid are substituted by the microcrystalline particles, and the atomic form factor is replaced by the independent scattering function of the particles. The equations have to be further modified into a two-dimensional form in connection with the cylindrical symmetry of the particles. Assuming that all particles are of one kind and denoting the radial distance of the scattering particles from the origin by r and the polar angle of this direction in the equatorial plane by α, the following formulas are obtained:

The coherent part of the amplitude scattered by the nth particle with a center at r_n is

$$F_n=\phi(kR)\exp(-ikr_n\cos\alpha) \tag{2}$$

and the scattered intensity:

$$I=\sum_m \sum_n F_m F_n J_0(kr_{mn}). \tag{3}$$

Following the usual procedure this double summation may be replaced by an integral if a continuous distribution of scattering cylinders exists, and the following final equation is obtained for the scattered intensity:

$$I=N\phi^2(kR)\left[1+\int_0^\infty 2\pi r(\rho(r)-\rho_0)J_0(kr)dr\right]. \tag{4}$$

[9] F. Zernike and J. A. Prins, Z. Physik **41**, 184–94 (1927); J. A. Prins, Naturwiss. **21**, 435–42 (1931); Z. Physik **56**, 617–48 (1929).

[10] P. Deybe, Ann. phys. **46**, 809–23 (1915); Physik. Z. **31**, 419–28 (1930); P. Debye and H. Menke, Physik. Z. **31**, 797–798 (1930); Ergeb. tech. Röntgenk. **2**, 1 (1931).

[11] S. Gingrich, Revs. Modern Phys. **15**, 90–110 (1943).

[12] G. Oster and D. P. Riley, Acta Cryst. **5**, 1–6 and 273–276 (1952).

In this equation $\rho(r)$ is the average density of scattering material as a function of the distance r from the origin cylinder and ρ_0 the average over-all density. The variation of $\rho(r)$ with distance can be expressed by a radial distribution function of scattering material in analogy with the distribution function for liquids (Zernike and Prins, and Prins[9]). For the present purpose it suffices to note that this function will in general be represented by a fluctuating curve which converges towards a constant density value, ρ_0. According to Eq. (4) the contribution from interparticle interference depends on the shape of this function. If a certain periodicity exists in the distribution of particles, the function will have pronounced maxima and minima at certain values of r and the amplitude scattered by the particles around each single particle will play an important part. If a random distribution of particles exists, the distribution function will be a straight line indicating a constant average density: $\rho(r)=\rho_0$. The part under the integral in Eq. (4) will then be zero which means that the scattering from all particles surrounding each individual particle balances out. The total radiation will in that case equal the sum $N\phi^2(kR)$ of the intensities from each separate particle as discussed under "*independent scattering.*"

The point of interest here is that a transition from the first type of distribution function into the second one will take place if the particles are moved further apart, for instance by swelling. This will have two results. In the first place, the over-all density of scattering material will be lower and consequently the maxima and minima of the distribution curve will be decreased. In the second place, a more irregular distribution of particles will result. The distribution curves from different points and in different directions will then no longer be in conformity with each other and counterbalance. Thus, with increased swelling the over-all distribution curve will flatten out and rapidly approach a straight line.

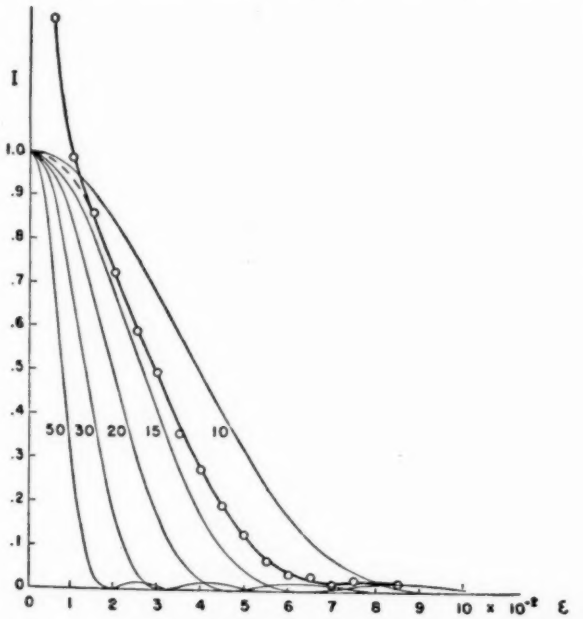


FIG. 2. A family of theoretical scattering curves for different particle radii. Ordinate: intensity 1. Abscissa: scattering angle in radians. The thick line with points marked on it represents an experimental scattering curve to be discussed in Sec. 2 of the experimental part.

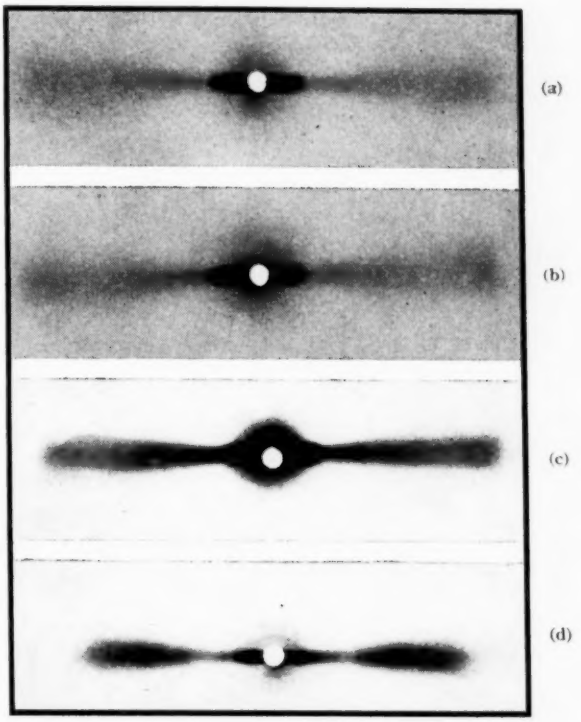


FIG. 3. Complete equatorial x-ray scattering by jute; original size. Sample distance 10 centimeters, pin-hole method. (a), (b), and (c), copper radiation; (d), chromium radiation. (a) 2-hour exposure. (b) 4-hour exposure. (c) 12-hour exposure. (d) 4-hour exposure—chromium radiation.

From the foregoing, the important conclusion follows that the contribution from interparticle interference can, in theory, be eliminated by subjecting the fiber to a sufficient degree of swelling. In the experimental part, it will be investigated whether and under which conditions it is actually possible to fully eliminate interparticle interference by following this principle.

EXPERIMENTAL PART

1. X-Ray Technique

The scattering experiments at small angles were carried out with a highly monochromatized focused beam obtained with two-combined quartz crystals. The first crystal is of the Johansson type (bent to a radius of 250 mm); the second crystal is bent to a larger radius of curvature. The crystals are mounted in such a way that their axes of curvature are parallel, and that the focal spot of the first crystal is the source of origin for the second crystal (see Guinier *et al.*[13]). Such line

[13] A. Guinier and G. Fournet, Compt. rend. **226**, 565 (1948); A. Guinier, Compt. rend. **223**, 31–32 (1946).

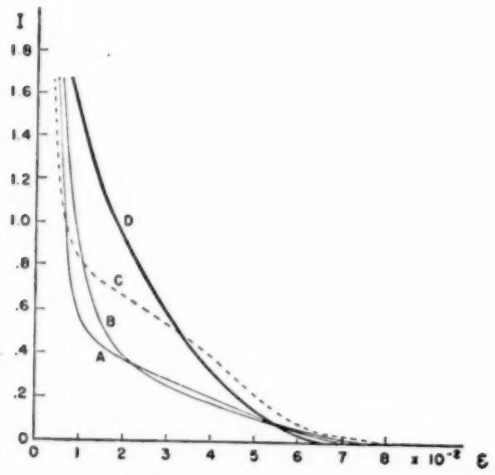


FIG. 4. Distribution of intensity in small-angle scattering by jute. A. Fiber in the dry state. B and C. Fiber in the water swollen state. D. Fiber swollen in 5 percent sodium hydroxide. Ordinate: density which is linear proportional to intensity. Abscissa: scattering angle in radians.

focusing suffices for the study of this material of cylindrical symmetry.

The measurement of intensity was carried out photographically; the density was measured with a densitometer equipped with a pin hole of 0.008-inch diameter. In calibration experiments the density was found to be linearly proportional to the exposure.

2. Equatorial Scattering by a Cellulose Fiber

As an illustration of the equatorial scattering by a natural cellulose fiber, the complete scattering by jute is illustrated in the photographs of Fig. 3. These pictures were obtained with a pin-hole beam at a sample distance of 10 cm. The scattering consists of a narrow band extending along the equator and of a width of only little more than the width of the x-ray beam at the film. The intensity rapidly decreases with scattering angle and reaches a definite minimum. A very important feature is that this minimum does not noticeably shift with longer exposures once the minimum exposure necessary for fully recording the scattering has been reached. This is illustrated by Fig. 3 in which the scattering by jute at increasing exposures of 2, 4, and 12 hours is given, and in which the angular location of the minimum is constant. The average values of the angular location of the minimum obtained with 4 different samples of pin-hole exposures of 10 centimeters sample distance is 0.072 radian corresponding to $R=14$ A.

Beyond this minimum a weak equatorial scattering extends up to the innermost diffraction spots. This scattering at wider angles is much weaker than the scattering at small angles. The intensity in this area gradually reaches a faint maximum. This part of the scattering is probably complicated. Besides small particle scattering, it may include scattering from the molecules of the amorphous regions, and will not be further analyzed here.

3. Independent Particle Scattering by a Cellulose Fiber; Direct Evaluation of Scattering

As an illustration of scattering at small angles, the curves obtained with jute fibers will be discussed. Figure 4 gives the scattering curves of this fiber at different degrees of swelling, dry, (A), swollen in water, (B and C), and swollen in 5 percent sodium hydroxide, (D). It may be seen from the curves that swelling causes a fargoing change in the over-all intensity and its distribution. The over-all intensity increases more and more with swelling, and the distribution curve approaches more and more a bell-shaped (Gaussian) curve. When the swelling is increased above a certain degree (in the case of jute, 5 percent sodium hydroxide), the scattering curve does no longer undergo further increase, and a final stage [curve (D)] seems to be reached. At higher swelling a slight recession from this final curve may even be found.

With some fibers, like certain rayons, the curves at lower swelling may show a distinct maximum and minimum (Fig. 5). This, as well as the differences between the curves (A), (B) and (C) from (D) in the case of jute, has been explained as a result of interparticle interference.[1]

It will be investigated in how far the curve (D) of jute (Fig. 6), corresponds to the theoretical curve for independent scattering. According to theory (see Fig. 1), the first minimum should be reached at a value of $kR=3.8$. From the angular location of this minimum at 0.072 radian, observed in the foregoing section, it can be calculated that the angular distance of the half intensity value (at $kr=1.61$) should be located at an

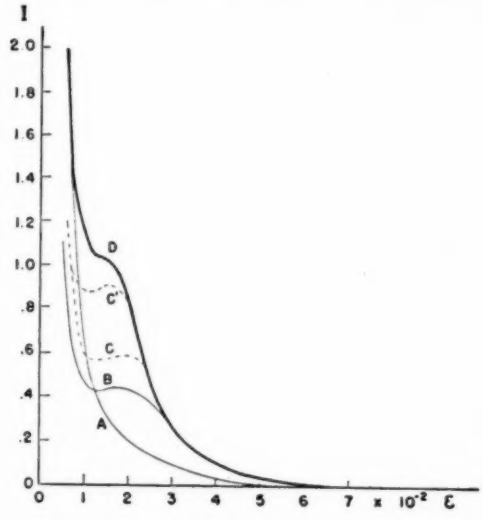


FIG. 5. Distribution of intensity in small-angle scattering by "Fortisan" rayon. A. Fiber in the dry state. B, C, and C'. Fiber swollen in water. D. Fiber swollen in 5 percent sodium hydroxide and stretched. Ordinate: density which is linear proportional to intensity. Abscissa: scattering angle in radians (from A. N. J. Heyn, 1953).

angular distance of 0.033 radian. Taking the intensity at 0.033 radian as 50 percent of the intensity at zero angle, and converting all other intensities accordingly, the complete scattering curve can be normalized for intensity. A comparison of this curve with the theoretical curves is made in Fig. 2. The normalized curve is found to be located between the curves for 15 and 20 A and to closely correspond again to a theoretical curve for a particle radius of 14 A. A direct comparison with the general theoretical curve of Fig. 1 is possible by plotting the experimental curve as a function of kR. This is done in Fig. 7. The experimental curve is represented by the points; the continuous line in the same figure is the theoretical curve. The agreement of experimental and theoretical curves is very good. The part at very small angles cannot be compared because of the finite width of the primary beam.

It follows from these observations, that under correct conditions of swelling, simple independent particle scattering can be actually realized, unobscured by phenomena like interparticle interference and multiple scattering which would result in more complicated scattering curves. (For the theory of multiple scattering by a system of parallel cylinders, refer, for example, to Twersky.[14]) It also follows that the microcrystallites may be actually treated as parallel cylinders of infinite length and sufficiently uniform radius.

Since the mathematical expression for the scattering function is known, it will be possible to calculate the particle size from the observed angular distribution of intensity. In Sec. 5 the conditions for independent scattering will be further analyzed.

4. Approximation of Scattering Function by Exponential Function

Instead of following the above direct method of curve matching, the particle size can also be determined

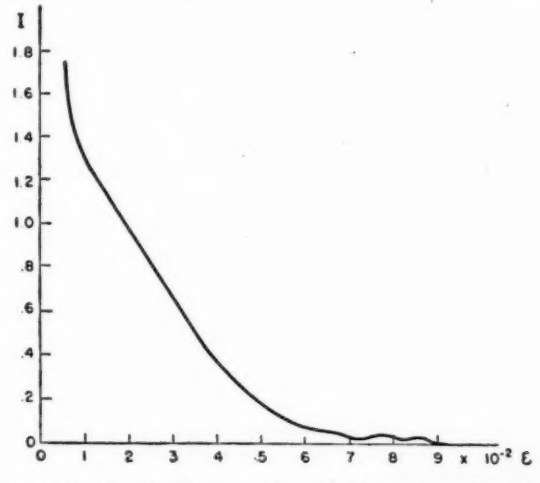


FIG. 6. Angular distribution of intensity in small-angle scattering by jute swollen in 5 percent sodium hydroxide. Ordinate: density which is linear proportional to the intensity. Abscissa: scattering angle in radians.

[14] Vic Twersky, J. Acoust. Soc. Am. **24**, 42–49 (1952).

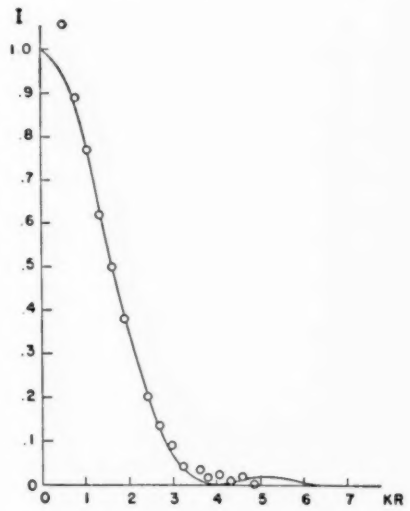


FIG. 7. Comparison of experimental and theoretical scattering in jute. The experimental curve is converted for standardized intensity and as a function of kR. The points are the experimental observations; the continuous line is the theoretical curve and is identical with the curve of Fig. 1.

indirectly from the decline of intensity of the independent scattering curve. This can be done following the method of Guinier[2] in which the scattering formula is approximated by a suitable exponential function. Since the small particle scattering was found to be expressed by the function $[2J_1(kR)/kR]^2$, the exponential approximating function to be used here should be $\exp(-\pi^2R^2\epsilon^2/\lambda^2)$ in which R represents the radius of the cylinder. The radius, R, is obtained following Guinier, from the slope, $-\pi^2R^2/\lambda^2 \log e$, of a plot of $\log I$ *versus* ϵ^2.

A straight logarithmic plot can only be expected to be found as far as the approximation formula actually approximates the Bessel function (which is nearly so) and if the scattering is actually expressed by a Bessel function. It follows from the theoretical discussion that the latter condition will only be fulfilled if the experimenter succeeds in obtaining independent particle scattering. A straight logarithmic plot can thus conversely be considered an indication for the realization of independent particle scattering.

If the logarithmic plot found is *not* a straight line, it will not be warranted in general for use in the calculation of particle size, since it is very probable that in this case interparticle interference is not fully eliminated. This conclusion follows from the theory, but will also be directly verified in the following section. It will be shown that with the same material either straight or curved logarithmic plots can be obtained depending on the degree of swelling and hence, on the extent to which interparticle interference plays a part.

Another reason for curved logarithmic plots to result *might* be the nonuniformity of particle size. This condition, however, will probably have little influence in most fibers.† The introduction of a factor for particle-size

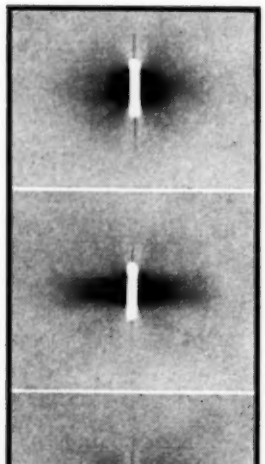


FIG. 8. Small-angle x-ray scattering by jute; original size, distance from sample to film, 20 cm. (a) Fiber in the dry state. (b) Fiber swollen in water. (c) Fiber swollen in 5 percent sodium hydroxide. (d) Fiber swollen in 10 percent sodium hydroxide.

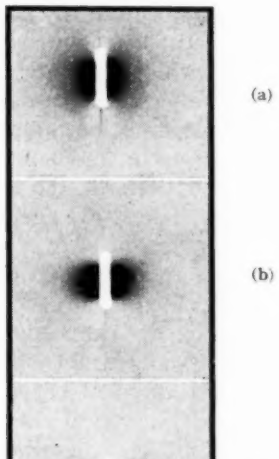


FIG. 9. Small-angle x-ray scattering by cotton; original size, distance from sample to film, 20 cm. (a) Fiber in the dry state. (b) Fiber swollen in water. (c) Fiber swollen in 10 percent sodium hydroxide. (d) Fiber swollen in 30 percent triton B.

distribution into the scattering equation will, namely, only slightly alter the basic shape of the curve and, thus, hardly influence the corresponding logarithmic plot.

The presence of interparticle interference should, therefore, be considered as the most probable reason for nonlinear plots, and it is certainly not allowed to ascribe such plots without any further consideration to nonuniformity of particles.[15]

5. Scattering by Fibers Under Different Conditions of Swelling; Analysis of Swelling Technique

Different fibers were investigated with the swelling method described in the foregoing section. Figures 8–10 give some photographs obtained in this way with jute, cotton, and ramie. In the dry state a narrow scattering band is only present in all cases; this band extends over a wider range along the equator when swelling is applied by mounting a fiber in water or various concentrations of alkalies or other swelling agents. The scattering curves of jute under different conditions of swelling are in Fig. 4. The discussion given in Sec. 3 may suffice to describe the differences in distribution of intensity at different degrees of swelling.

[15] M. H. Jellinek and I. Fankuchen, Ind. Eng. Chem. **17**, 158–64 (1945); (with E. Solomon) Ind. Eng. Chem. Anal. Ed. **18**, 172–75 (1946).

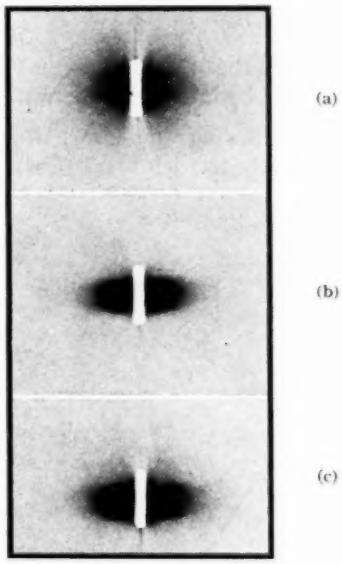


FIG. 10. Small-angle x-ray scattering by ramie; original size, distance from sample to film, 20 cm. (a) Fiber in the dry state. (b) Fiber swollen in 5 percent sodium hydroxide. (c) Fiber swollen in 10 percent sodium hydroxide.

In the present section, the logarithmic plots obtained following the method of Sec. 4 are studied instead of the original curves. In Fig. 11 the logarithmic plot from

the scattering by jute swollen in 5 percent sodium hydroxide is given. It may be seen that this curve is a completely straight line. In the next figures, plots for 3 different conditions of swelling are given in the same figure: in Fig. 12 for jute, in Fig. 13 for ramie, and in Fig. 14 for Cordura rayon.

It is seen from these figures that in all three cases a straight line is obtained under a certain condition of swelling. The condition under which a straight plot is obtained varies with the type of fiber investigated. With most fibers it is necessary to swell beyond the water-swollen condition. In the water-swollen condition, all fibers investigated yield curved plots (C).

If jute is swollen in 5 percent sodium hydroxide a straight logarithmic plot (B) is obtained, but if the same fiber is swollen in a higher concentration, namely in 10 percent sodium hydroxide, a curved plot (A) results again. With flax and ramie (Fig. 13), the situation is in so far different that these fibers yield already straight plots when swollen in 1 percent sodium hydroxide (B). If they are swollen in 5 percent sodium hydroxide, curved plots result again (A). The results with Cordura (Fig. 14), are even more interesting. In water or 10 percent sodium hydroxide a curved plot results; in 5 percent sodium hydroxide, the plot is also curved (A), but if swelling in 5 percent sodium hydroxide is combined with stretch, a completely straight plot is obtained (B).

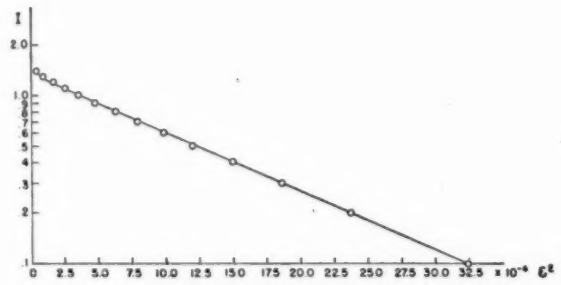


FIG. 11. Jute swollen in 5 percent sodium hydroxide. Log intensity plotted against square of scattering angle. Note perfect straight line obtained. Ordinate: log intensity; abscissa: square of scattering angle.

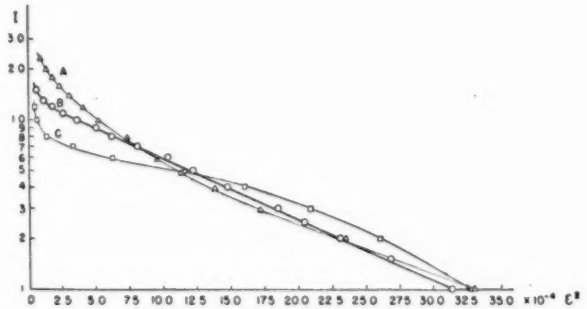


FIG. 12. Jute, log intensity plotted against square of scattering angle. A. Fiber swollen in 10 percent sodium hydroxide. B. Fiber swollen in 5 percent sodium hydroxide (a straight plot is obtained). C. Fiber swollen in water. Ordinate: log intensity; abscissa: square of scattering angle.

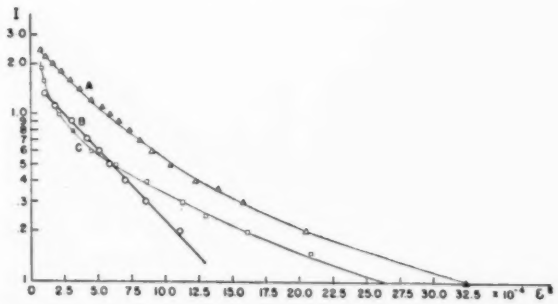


FIG. 13. Ramie log intensity plotted against square of scattering angle. A. Fiber swollen in 5 percent sodium hydroxide. B. Fiber swollen in 1 percent sodium hydroxide. C. Fiber swollen in water. Ordinate: log intensity; abscissa: square of scattering angle.

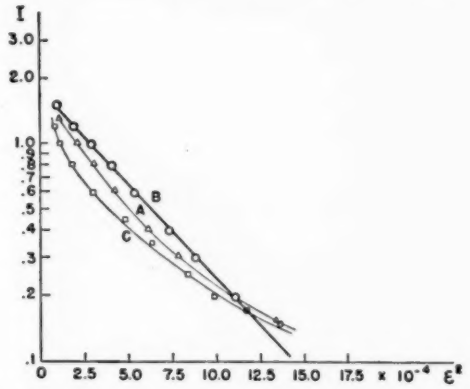
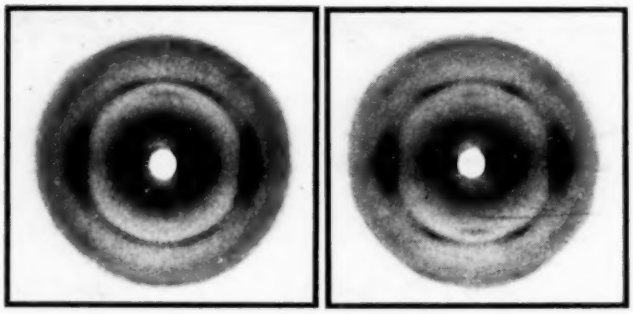


FIG. 14. Cordura rayon—log intensity plotted against square of scattering angle. A. Fiber swollen in 5 percent sodium hydroxide. B. Fiber swollen in 5 percent sodium hydroxide and stretched during the exposure (a straight plot is obtained). C. Fiber swollen in water. Ordinate: log intensity; abscissa: square of scattering angle.

These remarkable results can be explained in the following way. A straight plot will be obtained only if interparticle interference is eliminated by separating the particle sufficiently. Therefore, no straight plots are obtained below a certain degree of swelling, but at higher swelling, in higher concentrations of the swelling agent, a range exists in which interparticle interference is eliminated so that straight curves result.

That curved plots result again if the swelling is further increased (in still higher concentration of the agent) can be explained by the disorientation or rotation of the crystallites (from the original parallel position), which goes hand in hand with swelling. That such rotation actually takes place is seen from Fig. 15. Figure 15(a) is a diffraction photograph of Cordura swollen in 5 percent sodium hydroxide. This corresponds to the case of Fig. 14(A). From the size of the arcs it may be concluded that parallel orientation of microcrystallites as found in the dry sample is not present. Figure 15(b) is a diffraction photograph of the same fiber swollen in 5 percent sodium hydroxide, but irradiated in stretched condition, corresponding to the case of Fig. 14(B). It may be seen that the arcs are much shorter indicating that the microcrystallites have a much better parallel orientation.



(a) (b)

FIG. 15. Diffraction patterns of Cordura swollen in 5 percent sodium hydroxide. Sample distance, 4 cm. (a) Fiber swollen and unstretched. (b) Fiber stretched and swollen.

If the parallel orientation of microcrystallites is disrupted, the conditions for the application of the scattering equation derived for independent scattering are no longer fulfilled. In the first place, the cylindrical particles are no longer parallel so that the scattering can no longer be expressed by a simple Bessel function, but will be extremely complicated. In addition, interparticle interference, between certain parts of neighboring particles, may be introduced as a result of the disorientation, while other parts of these particles are moved excessively apart. Under these conditions, a straight logarithmic plot, as obtained from the simple scattering function, can no longer be expected to result.

The experimental results with Cordura are in good agreement with this explanation. At lower concentrations of swelling agents, the microcrystallites are not far enough separated to obtain independent scattering, at higher concentration (5 percent alkali) the crystallites are rotated to such extent that the above results of disorientation will occur. If, however, this excessive rotation is prevented by application of stretch, a straight logarithmic plot is found in accordance with theory.

From the above experimental results, it follows that one is by no means allowed to assume that with higher swelling the conditions for independent particle scattering will be increasingly better realized; only a certain range of swelling offers the correct condition; and it may even be necessary in some instances to apply special measures for maintaining parallel orientation.

6. Particle Sizes Determined

On the basis of the results obtained in the foregoing, the particle size was determined in various fibers, investigated under appropriate swelling conditions following the methods of Secs. 4 and 5.

The experimental error in the measurement of a single exposure was determined by measuring the scattering diagram of a single jute sample four times, after which the logarithmic plots were made and the particle size calculated. The results are given in Table I.

For determining the error between different exposures, four different scattering pictures were taken in which the particle size was determined. These data are given in Table II. It can be seen from the table that the variability of the results is very small in both cases and averages not more than one A unit.

TABLE I. Average diameters of microcrystallites in various fibers from small-angle scattering.

Fiber	Diameter, A
jute	28
flax	28
ramie	43
cotton	55
Super Cordura	36
Fiber G	40
Cordura	43

TABLE II. Average diameter, A, of microcrystallites in jute (cellulose fiber).

Sample No.	Measurements with 4 different samples	Measurements with one diagram (sample No. 1)	
1	28.00	*a*	28.32
2	27.02	*b*	28.48
3	28.63	*c*	27.16
4	*29.38*	*d*	*27.74*
Average	*28*		*28*

The results of particle-size determination with different fibers are given in Table I. In connection with the small experimental error established above, it may be concluded that different fibers exhibit significant differences in particle size.

It should be noted that the data calculated are a measure for the radius of gyration if the particles do not have exact cylindrical shape.

ACKNOWLEDGMENT

The author is indebted to the Claude Kress Foundation of Clemson College, which purchased the x-ray apparatus and some additional equipment used in this study.

Čerenkov Radiation of Electrons Moving Parallel to a Dielectric Boundary

J. G. LINHART
Research Laboratories of the British Thomson-Houston Company, Ltd., Rugby, England
(Received July 16, 1954)

The phenomenon of Čerenkov radiations of electrons moving parallel to a dielectric boundary is investigated, using an eigenfunction expansion of the electromagnetic field. It is shown that the Čerenkov effect is similar to the excitation of surface waves in traveling wave tubes.

1. INTRODUCTION

WE propose to calculate the radiation emitted by a point charge in uniform motion on a straight line parallel to the plane face of a semi-infinite dielectric slab. Using the eigenfunction expansion of an electromagnetic field, it is shown that such mechanism may produce radiation and that the result may be interpreted as an excitation of surface waves.

The solution is divided into three parts as follows:

(a) The eigenfunctions and the density of the corresponding eigenvalues in the wave-number space are determined for the given boundary conditions (here a cubical cavity half-filled with dielectric).

(b) The equation for the excitation coordinate $q_\lambda(t)$ (the amplitude) of the λth eigenfunction is set up and solved for the prescribed motion of the point charge e.[1]

(c) The elementary q_λ solutions are integrated in order to obtain the radiation output.

If the radiation output is found to be different from zero, the contention that electrons in uniform rectilinear motion close to a dielectric boundary radiate is proven.

The elementary solution $q_\lambda(t)$ obtained for the point charge e may be regarded as a Green's function of the problem.[2] Thus the calculation of the radiation field and output of a system of charges or even extended streams does not involve any new principles and these are obtained by summing up the contributions from the elements of the stream.

The formulas obtained are in agreement with the formulas for Čerenkov radiation of electrons in an infinite and homogenous medium.

2A.

Let us consider the solutions of

$$\Delta\bar{A}+\epsilon\mu\frac{v^2}{c^2}\cdot\bar{A}=0 \qquad (1)$$

($\bar{A}$ = vector potential)

for a cubical cavity with perfectly conducting walls half-filled with dielectric of permittivity ϵ, the other half being evacuated (see Fig. 1).

In rectangular coordinates (xyz) the above Helmholtz vector equation may be separated into three component scalar equations:

$$\Delta A_x+\epsilon\mu\frac{v^2}{c^2}A_x=0, \qquad (1a)$$

$$\Delta A_y+\epsilon\mu\frac{v^2}{c^2}A_y=0, \qquad (1b)$$

$$\Delta A_z+\epsilon\mu\frac{v^2}{c^2}A_z=0. \qquad (1c)$$

Also since the field has no sources:

$$\mathrm{div}\bar{A}=0. \qquad (2)$$

The boundary conditions for $\bar{A}$ are

(i) $$A_{11}=0$$

(where A_{11} is the component of the vector $\bar{A}$ parallel to the walls of the cubical cavity) and

(iia) $$A_i\left(z=\frac{l}{2}+\Delta\right)=A_i\left(z=\frac{l}{2}-\Delta\right),$$

(iib) $$\left[\mathrm{curl}\bar{A}\left(z=\frac{l}{2}+\Delta\right)\right]_i=\left[\mathrm{curl}\bar{A}\left(z=\frac{l}{2}-\Delta\right)\right]_i,$$

where $\Delta\to 0$ and $i=x, y$.

It is well known that an elementary solution of Eqs. (1a, b, c) is

$$A_i=a_i\cdot\exp[j(k_1x+k_2y+k_3z)] \quad \text{where} \quad i=x, y, z, \qquad (3)$$

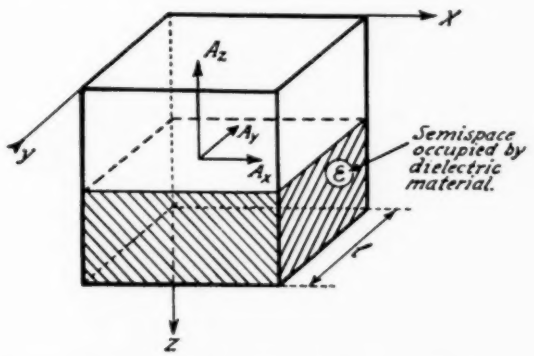


FIG. 1. The cubical cavity half-filled with a dielectric material.

[1] W. Heitler, *The Quantum Theory of Radiation* (Oxford University Press, London, England, 1944), p. 49.
[2] J. G. Linhart, J. Franklin Inst. **258**, No. 2 (1954).

and where k_1, k_2, and k_3 satisfy the characteristic equation

$$k_1^2+k_2^2+k_3^2=\epsilon\mu\frac{\nu^2}{c^2}, \tag{4}$$

which is an equation of a sphere of radius $R=(\epsilon\mu)^{\frac{1}{2}}(\nu/c)$ in the $k_1k_2k_3$ space.

Because the problem is homogeneous in the x and y directions the solutions for A_z (satisfying the appropriate boundary conditions) are simple harmonic functions in x and y. Because of this and also because of the divergence equation (2) the A_x and A_y must also be simple harmonic functions of x and y; k_1 and k_2 must therefore be real numbers, having the same values in both halves of the cavity. Thus it follows from Eq. (4) that in the upper half:

$$k_3=\text{is purely real}=\gamma,$$

or

$$k_3=\text{is purely imaginary}=i\beta.$$

Using the boundary conditions i it follows that:

$$\begin{aligned} A_x&=a_x\cdot\cos k_1x\cdot\sin k_2y\cdot\sin k_3(z-l),\\ A_y&=a_y\cdot\sin k_1x\cdot\cos k_2y\cdot\sin k_3(z-l),\\ A_z&=a_z\cdot\sin k_1x\cdot\sin k_2y\cdot\cos k_3(z-l), \end{aligned} \tag{3a}$$

which represent the field in the dielectric-filled lower half of the cube. Similarly

$$\begin{aligned} A_x^1&=a_x^1\cdot\cos k_1x\cdot\sin k_2y\cdot\frac{\sin\gamma z}{\sinh\beta z},\\ A_y^1&=a_y^1\cdot\sin k_1x\cdot\cos k_2y\cdot\frac{\sin\gamma z}{\sinh\beta z},\\ A_z^1&=a_z^1\cdot\sin k_1x\cdot\sin k_2y\frac{\cos\gamma z}{\cosh\beta z}, \end{aligned} \tag{3b}$$

which represent the field in the evacuated upper half of the cube.

Let us treat the case when $k_3=i\beta$ since this gives a field capable of interaction with electrons of smaller velocities than the speed of light in vacuum (a field equivalent to the totally reflected field from an ∞ dielectric boundary).

The eigenvalues k_1 and k_2 are found from the boundary conditions (i):

$$k_1l=m_1\pi \quad\text{and}\quad k_2l=m_2\pi,$$

where $m_{1,2}=1, 2\cdots n$,

$$k_1=\frac{m_1\pi}{l};\quad k_2=\frac{m_2\pi}{l}.$$

Let us eliminate k_1 and k_2 using Eq. (4) written for both halves of the cube:

$$k_1^2+k_2^2+k_3^2=\frac{\nu^2}{c^2}\epsilon \text{ for the dielectric (2)}, \tag{4a}$$

$$k_1^2+k_2^2-\beta^2=\frac{\nu^2}{c^2} \text{ for vacuum (1)}. \tag{4b}$$

Subtracting we get:

$$k_3^2+\beta^2=\frac{\nu^2}{c^2}(\epsilon-1), \tag{5}$$

which is the equation of a circle of radius $\rho_0=(\nu/c)(\epsilon-1)^{\frac{1}{2}}$.

The waves whose $k_3<\rho_0$ (i.e. those for which $\beta=$ real) suffer a total reflection at the dielectric boundary, whereas those for which $\epsilon(\nu/c)>k_3>\rho_0$ are partly refracted [the $\frac{\sin}{\cos}\gamma z$ functions in Eqs. (3b)] and partly reflected.

It is easy to show that only the fields for which $k_3<\rho_0$ may interact with charged particles (whose velocity $v<c$).

Since the phase of the field must remain constant for an interaction to take place we have for electrons moving parallel to the dielectric boundary,

$$(k_1^2+k_2^2)^{\frac{1}{2}}\cdot x-\nu t=\text{const.}$$

Thus,

$$\frac{\partial x}{\partial t}=v=\frac{\nu}{(k_1^2+k_2^2)^{\frac{1}{2}}}, \tag{6}$$

but, from Eq. (4a),

$$k_1^2+k_2^2=\frac{\nu^2}{c^2}\epsilon-k_3^2.$$

The maximum k_3 is obtained by taking $v=c$. Then from Eq. (6)

$$\begin{aligned} c^2&=\frac{\nu^2}{(\nu^2/c^2)\epsilon-k_3^2},\\ k_3&=\frac{\nu}{c}(\epsilon-1)^{\frac{1}{2}}=\rho_0. \end{aligned} \tag{7}$$

The division of the $\bar{k}$-space into two regions by the cone $k_3=\rho_0$ is depicted in Fig. 2.

When the dimensions of the cubical cavity

$$l\to\infty,$$

the eigenvalues ν form a continuous spectrum and it follows from Eq. (4a) that k_3 is of the same form as k_1 and k_2, i.e.:

$$k_3=m_3\pi/l.$$

Thus when the transition to an infinite enclosure is made, the $\bar{k}$-space is filled uniformly with eigenvectors

$$\bar{k}=\bar{\alpha}_1k_1+\bar{\alpha}_2k_2+\bar{\alpha}_3k_3,$$

where $\bar{\alpha}_1\bar{\alpha}_2\bar{\alpha}_3$ are unit vectors in the $k_1k_2k_3$ space, and

$$k_i = m_i\pi/l, \quad i=1, 2, 3,$$

where $m_1m_2m_3$ are positive integers.

The spacing Δk between two adjacent eigenvalues is (see Fig. 3)

$$\Delta k = \pi/l.$$

Thus the density of eigenvalues in the $\bar{k}$-space is

$$f = l^3/\pi^3. \tag{8}$$

The wave numbers k_1, k_2, and k_3 may be expressed in terms of the angle of incidence θ and the angle Ψ between the plane of incidence and the axis x (*viz.* Fig. 4). We have:

$$k_1 = \epsilon^{\frac{1}{2}}\frac{\nu}{c}\cdot\sin\theta\cdot\cos\Psi, \quad k_2 = \epsilon^{\frac{1}{2}}\frac{\nu}{c}\sin\theta\cdot\sin\Psi, \tag{9a,b}$$

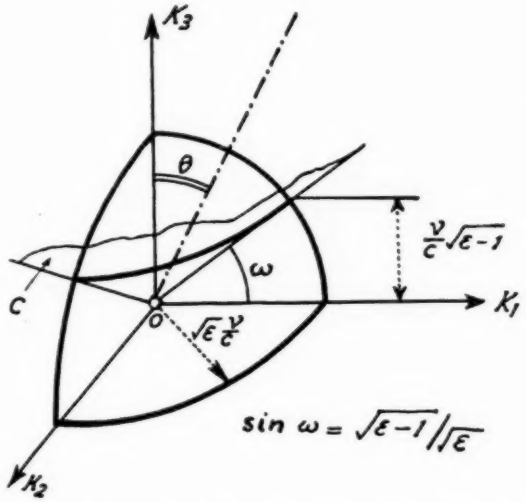


FIG. 2. The k-space divided into two parts by the cone C.

and therefore

$$k_1^2 + k_2^2 = (\nu^2/c^2)\epsilon\sin^2\theta.$$

Also

$$k_3^2 = (\nu^2/c^2)\epsilon\cdot\cos^2\theta, \tag{9c}$$

and from Eq. (4b) we get the corresponding β:

$$\beta = \frac{\nu}{c}[\epsilon\sin^2\theta - 1]^{\frac{1}{2}}. \tag{9b}$$

Using the latter expressions for k_1, k_2, k_3, and β, passing to the limit of $l\to\infty$, and making a transformation $-z' = z - l/z$ we may write the Eqs. (3a, b) in the following form:

$$A_x = a_x \begin{matrix}\cos\\ \sin\end{matrix} k_1x \begin{matrix}\sin\\ \cos\end{matrix} k_2y\cdot\sin(k_3z+\alpha),$$

$$A_y = a_y \begin{matrix}\sin\\ \cos\end{matrix} k_1x \begin{matrix}\cos\\ \sin\end{matrix} k_2y\cdot\sin(k_3z+\alpha), \tag{10a}$$

$$A_z = a_z \begin{matrix}\sin\\ \cos\end{matrix} k_1x \begin{matrix}\sin\\ \cos\end{matrix} k_2y\cdot\cos(k_3z+\alpha).$$

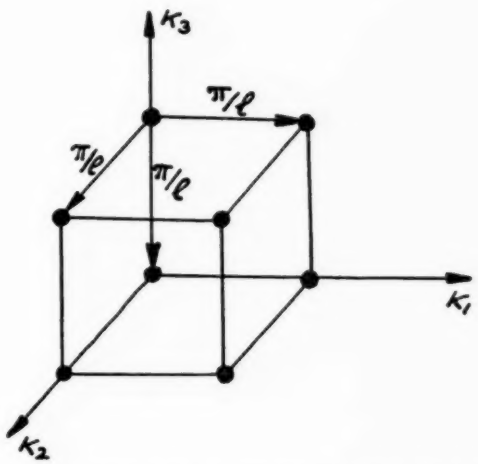


FIG. 3. The spacing of the eigenvector termini in the k-space.

$$A'_x = a'_x \begin{matrix}\cos\\ \sin\end{matrix} k_1x \begin{matrix}\sin\\ \cos\end{matrix} k_2y\cdot\exp\left[-\frac{\nu}{c}z(\epsilon\sin^2\theta-1)^{\frac{1}{2}}\right],$$

$$A'_y = a'_y \begin{matrix}\sin\\ \cos\end{matrix} k_1x \begin{matrix}\cos\\ \sin\end{matrix} k_2y\cdot\exp\left[-\frac{\nu}{c}z(\epsilon\sin^2\theta-1)^{\frac{1}{2}}\right], \tag{10b}$$

$$A'_z = a'_z \begin{matrix}\sin\\ \cos\end{matrix} k_1x \begin{matrix}\sin\\ \cos\end{matrix} k_2y\cdot\exp\left[-\frac{\nu}{c}z(\epsilon\sin^2\theta-1)^{\frac{1}{2}}\right].$$

It should be appreciated that to every eigenvector $k_1k_2k_3$ there corresponds two orthogonal eigenfields of mutually orthogonal planes of polarization. We shall choose the plane of incidence (p) and the dielectric boundary plane (π) to be the two basic vibration planes. The fields whose $\bar{A}$ vector lies in the plane p will be called the p-field, and the field whose $\bar{A}$ vector lies in the π plane will be the π-field. Thus it follows that (see Fig. 4):

$$(a_x^2+a_y^2)_p^{\frac{1}{2}} = a_p\cdot\cos\theta; \quad a_{zp} = a_p\cdot\sin\theta,$$
$$(a_x^2+a_y^2)_\pi^{\frac{1}{2}} = a_\pi; \quad a_{z\pi} = 0.$$

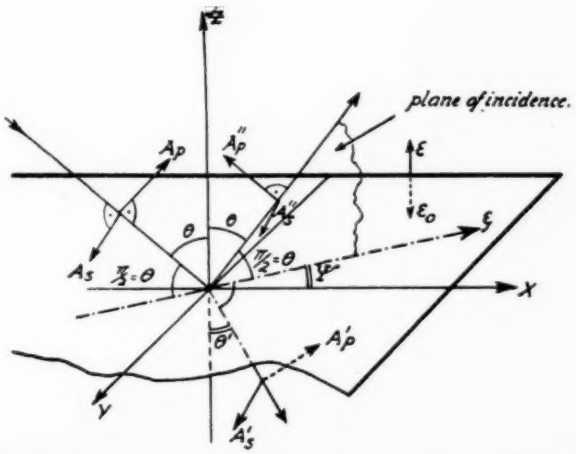


FIG. 4. The geometry of the reflection of electromagnetic waves at the dielectric boundary.

The magnitude of a_p and a_π follows from the normalizing condition (see reference 1, p. 42.)

$$\epsilon \int |\bar{A}|^2 dV = 4\pi c^2,$$

where $V = l^3$. Thus,

$$a = a_p = a_\pi = \left(\frac{8\pi c^2}{\epsilon l^3}\right)^{\frac{1}{2}}. \tag{11}$$

It now remains to find the volume of a'_p and a'_π. These may be obtained from the boundary conditions (iib) (or from Fresnel's formulas). The final form of Eqs. (10b) is:

$$\begin{aligned}
A'_{x\pi} &= -a\left(\frac{\epsilon}{\epsilon-1}\right)^{\frac{1}{2}} \cos\theta \cdot \sin\Psi \cdot \begin{matrix}\cos\\ \sin\end{matrix} k_1 x \begin{matrix}\sin\\ \cos\end{matrix} k_2 y e^{-\beta z}, \\
A'_{y\pi} &= a\left(\frac{\epsilon}{\epsilon-1}\right)^{\frac{1}{2}} \cos\theta \cdot \sin\Psi \begin{matrix}\sin\\ \cos\end{matrix} k_1 x \begin{matrix}\cos\\ \sin\end{matrix} k_2 y e^{-\beta z}, \\
A'_{z\pi} &= 0.
\end{aligned} \tag{12a}$$

$$\begin{aligned}
A'_{xp} &= -a\cos\Psi\cos\theta\left(\frac{\epsilon}{\epsilon-1}\right)^{\frac{1}{2}} \\
&\quad \times\left[\frac{\epsilon\sin^2\theta - 1}{(\epsilon+1)\sin^2\theta - 1}\right]^{\frac{1}{2}} \begin{matrix}\cos\\ \sin\end{matrix} k_1 x \begin{matrix}\sin\\ \cos\end{matrix} k_2 y \cdot e^{-\beta z}, \\
A'_{yp} &= -a\sin\Psi\cos\theta\left(\frac{\epsilon}{\epsilon-1}\right)^{\frac{1}{2}} \\
&\quad \times\left[\frac{\epsilon\sin^2\theta - 1}{(\epsilon+1)\sin^2\theta - 1}\right]^{\frac{1}{2}} \begin{matrix}\sin\\ \cos\end{matrix} k_1 x \begin{matrix}\cos\\ \sin\end{matrix} k_2 y \cdot e^{-\beta z}, \\
A'_{zp} &= a\frac{\epsilon}{(\epsilon-1)^{\frac{1}{2}}} \\
&\quad \times\frac{\sin\theta\cos\theta}{[(\epsilon+1)\sin^2\theta - 1]^{\frac{1}{2}}} \begin{matrix}\sin\\ \cos\end{matrix} k_1 x \begin{matrix}\sin\\ \cos\end{matrix} k_2 y \cdot e^{-\beta z}.
\end{aligned} \tag{12b}$$

These are the eigenfunctions we shall need in the next section.

2B.

Having obtained the eigenfunctions of the available space, we shall now find their excitation by solving the forced oscillation equation[1]:

$$\ddot{q} + \nu^2 q = \frac{e}{c}\bar{v}\cdot\bar{A}, \tag{13a}$$

where $q(t)$ is the time variation of the pattern $\bar{A}(xyz)$ and $\bar{v}$ is the velocity vector of the charge e. In our case $\bar{v}$ is a constant vector. Without loss of generality let us choose the trajectory to be $y=0$, $z=z_0$ and the position of the point charge (see Fig. 5)

$$x = x_0 + vt.$$

Thus the driving force represented by the right-hand side of Eq. (13a) becomes:

$$F_\nu = \frac{e}{c}\bar{v}\cdot\bar{A} = \frac{e}{c} v \cdot a'_x e^{-\beta z_0} \begin{matrix}\sin\\ \cos\end{matrix}\left\{\frac{\nu}{c}\epsilon^{\frac{1}{2}}\sin\theta\cos\Psi(x_0+vt) - \alpha\right\}.$$

Choosing $x_0 = 0$ for $t = 0$ we get:

$$\ddot{q} + \nu^2 q = \frac{e}{c} v \cdot a'_x \cdot e^{-\beta z_0} \begin{matrix}\sin\\ \cos\end{matrix}\left\{\frac{\nu}{c}\epsilon^{\frac{1}{2}}\sin\theta\cdot vt\cdot\cos\Psi - \alpha\right\}. \tag{13b}$$

The a'_x may be considered to belong either to the p-field or the π-field x-component.

The solution of this equation is

$$q = \frac{F_{\pi p}}{\nu^2 - \nu_0^2}\left\{\begin{matrix}\sin\\ \cos\end{matrix}(v_0 t - \alpha) - \frac{\nu_0}{\nu}\cdot\begin{matrix}\sin\\ \cos\end{matrix}(\nu t - \alpha)\right\}, \tag{14}$$

where

$$\begin{aligned}
F_\pi &= \frac{e}{c} v a'_{x\pi} e^{-\beta z_0} = \frac{e}{c} v a_\pi \left(\frac{\epsilon}{\epsilon-1}\right)^{\frac{1}{2}} \cos\theta \\
&\quad \cdot \exp\left[-\frac{\nu}{c}(\epsilon\cdot\sin^2\theta - 1)^{\frac{1}{2}}\cdot z_0\right]\sin\Psi,
\end{aligned} \tag{15a}$$

$$\begin{aligned}
F_p &= \frac{e}{c} v a'_{xp} e^{-\beta z_0} \\
&= \frac{e}{c} v a_p \left(\frac{\epsilon}{\epsilon-1}\right)^{\frac{1}{2}} \frac{(\epsilon\cdot\sin^2\theta - 1)^{\frac{1}{2}}}{[(\epsilon+1)\sin^2\theta - 1^{\frac{1}{2}}]} \cos\theta \\
&\quad \cdot \exp\left[-\frac{\nu}{c}(\epsilon\cdot\sin^2\theta - 1)^{\frac{1}{2}}\cdot z_0\right]\cos\Psi,
\end{aligned} \tag{15b}$$

and

$$a_p^2 = a_\pi^2 = \frac{8\pi c^2}{\epsilon l^3}, \quad \nu_0 = \frac{\nu}{c} v \epsilon^{\frac{1}{2}}\cdot\sin\theta\cos\Psi. \tag{16}$$

At this stage it is convenient to combine two elementary fields 1 and 2, whose $\bar{k}$ vectors differ only in the

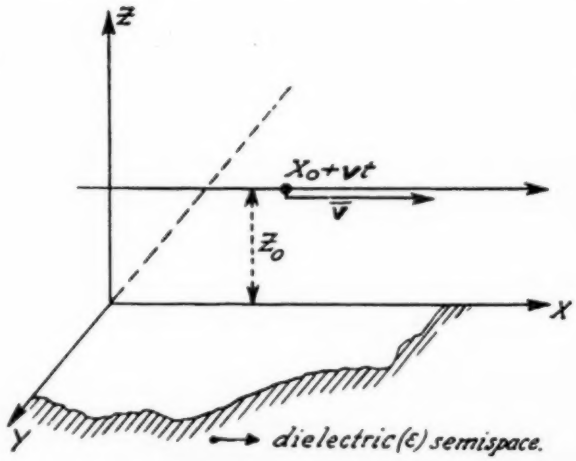


FIG. 5. The trajectory of the charge e.

angle Ψ, the latter satisfying a relation:

$$\Psi_1 = -\Psi_2.$$

In computing the $q' = q_1 + q_2$ of such a pair it follows that the $q_\tau = 0$, whereas

$$q'_p = 2q_p.$$

This indicates that only those waves will be excited whose plane of vibration is perpendicular to the dielectric boundary.

2C. THE FREQUENCY SPECTRUM

The energy output W_{ν_0} per unit angular frequency ($\Delta\nu_0 = 1$) radiated in the ν_0-radiation field is found using the formula for the radiation output in Appendix 1:

$$W_{\nu_0} = \int \dot{q}_\nu(t;\nu,\theta,\Psi) F_\nu(t;\nu,\Theta,\Psi)\cdot f \cdot d\Omega \text{ (erg/sec)}, \quad (17)$$

where f is given by Eq. (8), and Ω is a region in the k-space over which we integrate. This region comprises all the eigenvectors of the eigenfields that are excited (whose $q_\nu \neq 0$) by the moving charge e and are oscillating with angular frequency $\nu_0 \pm \frac{1}{2}$. Let us investigate more closely the integrand I in Eq. (17):

$$I_{1,2} = \frac{F_p{}^2 \cdot \nu_0}{\nu^2 - \nu_0{}^2}\left\{ \begin{matrix} \cos \\ -\sin \end{matrix} (\nu_0 t - \alpha) + \frac{\nu}{\nu} \begin{matrix} -\cos \\ \sin \end{matrix} (\nu t - \alpha) \right\} \cdot \begin{matrix} \sin \\ \cos \end{matrix} (\nu_0 t - \alpha)\cdot f. \quad (18)$$

It is clear that fields whose eigenvalue $\nu \approx \nu_0$ are excited most. Thus if $\nu \approx \nu_0$ the integrand I becomes

$$I_{1,2} = F_p{}^2 \begin{matrix} \sin^2 \\ \cos^2 \end{matrix} (\nu_0 t - \alpha) \frac{\sin(\nu - \nu_0)t}{2(\nu - \nu_0)} \cdot f.$$

Since we must add all the contributions to W_ν, the total integrand I is

$$I = I_1 + I_2 = \tfrac{1}{2} F_p{}^2 \frac{\sin(\nu - \nu_0)t}{\nu - \nu_0} \cdot f. \quad (19)$$

The region of integration Ω (see Fig. 6) is specified by considering only those eigenfields that oscillate with angular frequency $\nu_0 \pm \frac{1}{2}$. Since

$$\nu_0 = \nu - \frac{v}{c}\sqrt{\epsilon}\,\sin\theta\,\cos\Psi,$$

the region Ω is confined by two parallel planes

$$\nu_0 \pm \tfrac{1}{2} = \nu - \frac{v}{c}\sqrt{\epsilon}\,\sin\theta\,\cos\Psi.$$

We shall exhaust the region Ω by considering a series of annular segments for which

$$\sin\theta\,\cos\Psi = \text{const.}$$

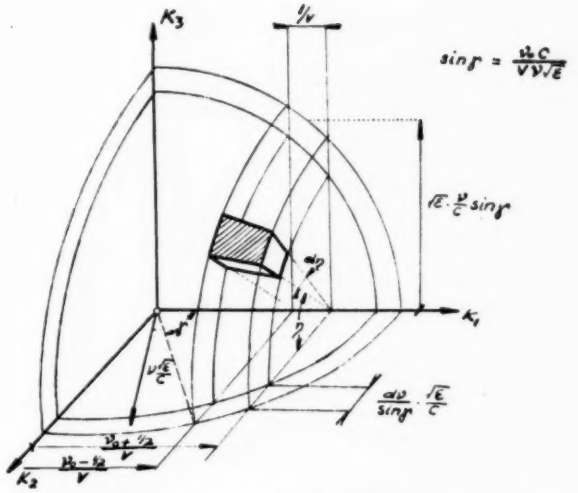


FIG. 6. The region of integration (in the k-space) relevant to the integral occurring in Eq. (17) and to Eq. (20).

The volume of such a segment is (see Fig. 6)

$$d\Omega = \sqrt{\epsilon}\,\frac{\nu}{c}\sin\gamma \cdot \frac{1}{v} \cdot \frac{d\nu(\epsilon^{\frac{1}{2}}/c)}{\sin\gamma} \cdot d\eta, \quad (20)$$

$$d\Omega = \frac{\epsilon}{vc^2}\nu d\nu d\eta.$$

Thus we have

$$W_{\nu_0} = \tfrac{1}{2}\int_{\nu(v_p/v)\,\nu_0}^{\infty}\int_{\eta=0}^{\eta=\pi/2} F_p{}^2[\nu,\theta(\eta),\Psi(\eta)] \times \frac{\sin(\nu-\nu_0)t}{\nu-\nu_0}\cdot f \cdot \frac{\epsilon}{vc^2}\nu d\nu d\eta.$$

Since $F^2{}_p \cdot \nu$ is a slowly varying function of ν and the greatest contribution to W_{ν_0} is expected to be near $\nu = \nu_0$, we may write

$$W_{\nu 0} = \frac{\epsilon\nu_0 f}{2vc^2}\int_{\eta=0}^{\pi/2} F_p{}^2[\nu_0,\theta(\eta),\Psi(\eta)] \cdot d\eta \cdot \int_{\nu_0(v_p/v)}^{\infty} \frac{\sin(\nu-\nu_0)t}{\nu-\nu_0} d\nu. \quad (20a)$$

The last integral yields

$$\int_{\nu_0(v_p/v)}^{\infty} \frac{\sin(\nu-\nu_0)t}{\nu-\nu_0} d\nu \cong \int_{-\infty}^{\infty} \frac{\sin x}{x} dx = \pi.$$

Thus,

$$W_{\nu 0} = \frac{\pi\epsilon\nu_0 f}{2vc^2}\int_{\eta=0}^{\pi/2} F_p{}^2(\nu_0,\eta) d\eta \text{ (ergs/sec)}. \quad (20b)$$

We now need relationships between λ (or rather the angle η) and the angles θ and Ψ. These follow easily

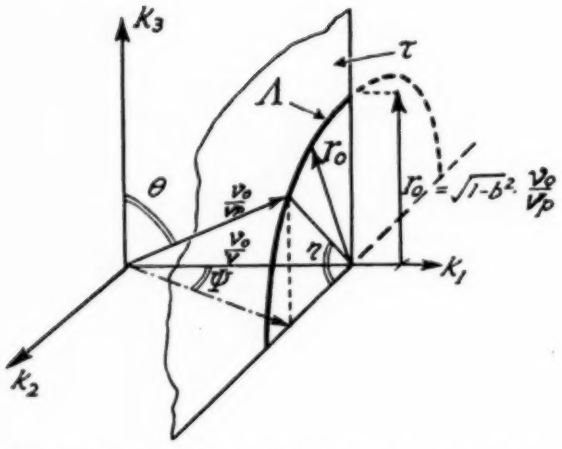


FIG. 7. The geometrical relationship between λ, θ, and Ψ.

from Fig. 7:

$$\cos\theta=(1-b^2)^{\frac{1}{2}}\cdot\sin\eta,$$

$$\sin\theta=[1-(1-b^2)\sin^2\eta]^{\frac{1}{2}},$$

$$\sin\Psi=(1-b^2)^{\frac{1}{2}}\cdot\cos\eta,$$

$$\cos\Psi=\left[1+\frac{1-b^2}{b^2}\sin^2\eta\right]^{\frac{1}{2}},$$

where $b=c/\sqrt{\epsilon}v=v_p/v$ and v_p= phase velocity in dielectric medium of dielectric constant ϵ.

The F_p^2 expressed in terms of η reads:

$$F_p^2=\frac{e^2}{c^2}v^2a_p^2\frac{\epsilon}{\epsilon-1}(1-b^2)\sin^2\eta\frac{(\epsilon-1)+\epsilon(1-b^2)\sin^2\eta}{\epsilon-(\epsilon+1)(1-b^2)\sin^2\eta}$$

$$\times\left[1+\frac{(1-b^2)}{b^2}(\sin^2\eta)\right]$$

$$\cdot\exp\left\{-2\frac{\nu_0}{c}[(\epsilon-1)-\epsilon(1-b^2)\sin^2\eta]^{\frac{1}{2}}\cdot Z_0\right\}. \quad (15b)$$

Substituting these into the integral (18) we get:

$$W_{\nu_0}=\frac{4}{\pi}\frac{e^2}{c^2}\nu_0v(1-b^2)\int_0^{\pi/2}\sin^2\eta$$

$$\times\frac{1-\epsilon\dfrac{1-b^2}{\epsilon-1}\sin^2\eta}{1-\dfrac{\epsilon+1}{\epsilon}(1-b^2)\sin^2\eta}\left(1+\frac{1-b^2}{b^2}\sin^2\eta\right)$$

$$\times\exp\left\{-2\frac{\nu_0}{c}[(\epsilon-1)-\epsilon(1-b^2)\sin^2\eta]^{\frac{1}{2}}\cdot Z_0d\eta\right\}. \quad (20c)$$

This integral is difficult to solve but progress can be made if we restrict ourselves to values of v such that

$$v=\frac{c}{\sqrt{\epsilon}}+\delta,$$

where $\delta\epsilon^{\frac{1}{2}}/c\ll1$. Then the exponent in the integral in Eq. (20c) is:

$$-2\frac{\nu_0}{c}z_0[(\epsilon-1)-\epsilon(1-b^2)\sin^2\eta]^{\frac{1}{2}}\cong-2(\epsilon-1)^{\frac{1}{2}}$$

$$\times\left(1-\frac{\epsilon}{\epsilon-1}\cdot\frac{\delta}{v_p}\sin^2\eta\right)\cdot\frac{\nu_0z_0}{c}.$$

Substituting this into Eq. (20c) we have:

$$W_{\nu_0}=\frac{4}{\pi}\frac{e^2}{c^2}\nu_0v\cdot\left(\frac{2\delta}{v_p}\right)\times\exp\left[-2\frac{\nu_0z_0}{c}(\epsilon-1)^{\frac{1}{2}}\right]$$

$$\times\int_0^{\pi/2}\sin^2\eta\frac{1-\dfrac{2\epsilon}{\epsilon-1}\dfrac{\delta}{v_p}\sin^2\eta}{1-\dfrac{\epsilon+1}{\epsilon}\dfrac{\delta}{v_p}\sin^2\eta}\left(1+2\frac{\delta}{v_p}\sin^2\eta\right)$$

$$\times\exp\left[+2\frac{\nu_0z_0}{c}\frac{\epsilon}{(\epsilon-1)^{\frac{1}{2}}}\frac{\delta}{v_p}\sin^2\eta\right]\cdot d\eta. \quad (20d)$$

If one also assumes that $\epsilon\gg1$, which is usually true, we get:

$$W_{\nu_0}=2\frac{e^2}{c^2}\nu_0\cdot\delta\cdot\exp\left[-2\frac{\nu_0z_0}{c}(\epsilon-1)^{\frac{1}{2}}\right]\text{ (erg/sec)}. \quad (21)$$

This formula must satisfy the following conditions

(1) $W_{\nu_0}\to0$ as $v\to v_p$.
(2) $W_{\nu_0}\to0$ as $\epsilon\to1$.

(3) When $z_0\to0$, the expression Eq. (21) must functionally converge to the formula for the frequency spectrum of the Čerenkov radiation in an infinite dielectric medium which is[3]

$$W_{\nu_0}=\frac{e^2}{c^2}\nu v\left(1-\frac{v_p^2}{v^2}\right)\text{ (ergs/sec)}. \quad (22)$$

It follows from inspection of Eq. (21) that the conditions (1) and (2) are obviously satisfied (since $\delta\to0$ as either $v\to v_p$ or $\epsilon\to1$).

In order to prove (3) let us put in Eq. (22)

$$v=v_p+\delta,\quad \delta\ll v.$$

Thus Eq. (22) becomes

$$W_{\nu_0}\cong2\cdot\frac{e^2}{c^2}\nu_0\delta. \quad (23)$$

When $z_0\to0$ we have from Eq. (21)

$$W_{\nu_0}\cong2\frac{e^2}{c^2}\nu_0\cdot\delta, \quad (24)$$

which compares favorably with Eq. (23).

[3] L. I. Schiff, *Quantum Mechanics* (McGraw-Hill Book Company, Inc., New York, 1949), p. 265.

The most interesting functional feature of Eq. (21) is its dependence on ν_0. This is represented by the curve W_{ν_0} in Fig. 8.

The frequency ν_m at which most of the energy is radiated is obtained by finding the maximum of W_{ν_0}. Thus from

$$\frac{\partial W_{\nu_0}}{\partial \nu_0}=0$$

we get

$$1\cong\nu_m\cdot\frac{2z_0}{c}(\epsilon-1)^{\frac{1}{2}}. \tag{25a}$$

Putting

$$\nu_m=2\pi c/\lambda_m,$$

where λ_m is the corresponding wavelength, one obtains:

$$\lambda_m\cong 4\pi(\epsilon-1)^{\frac{1}{2}}\cdot z_0. \tag{25b}$$

Similarly, maximizing with respect to ϵ, we find $\epsilon_m=2$.

However, we must keep in mind our assumption that

$$\epsilon\gg 1.$$

Thus we cannot trust too much the value found for ϵ_m. However, using the latter in Eq. (25b), it follows that

$$\lambda_m\simeq 4\pi z_0.$$

Thus if one desired to generate

$$\lambda_m=1 \text{ cm},$$

the optimum distance z_0 of the electrons from the dielectric would be $z_0\cong 0.8$ mm.

The total energy per second radiated by the point charge e is

$$W=\int_0^\infty W_{\nu_0}d\nu_0.$$

$$W=2\frac{e^2}{c^2}\delta\int_0^\infty \nu_0\cdot\exp\left[-\frac{\nu_0 z_0}{c}(\epsilon-1)^{\frac{1}{2}}\right]\cdot d\nu_0. \tag{26}$$

$$W\cong 2e^2\frac{\delta}{z_0^2}\frac{1}{\epsilon-1}\ (\text{ergs/sec}).$$

It is interesting to note that here the ultraviolet catastrophe does not threaten to make W infinite, since the cut-off frequency ν_m is here clearly defined by z_0.

Our assumption that $\epsilon\gg 1$, on the other hand, causes $W\to\infty$ as $\epsilon\to 1$. Equation (26) is accurate to more than 10 percent for $\epsilon=2.5$ and the accuracy increases as ϵ increases.

3. CONCLUSIONS

The Čerenkov effect investigated in the previous sections is thus shown to be equivalent to the excitation of surface waves, since such waves are represented by the set of eigenfunctions used and the excitation of these is shown to be of finite amplitude. Thus in principle there is no difference between a Čerenkov oscillator of the type discussed and a traveling wave tube where electrons excite waves supported by a slow wave structure.

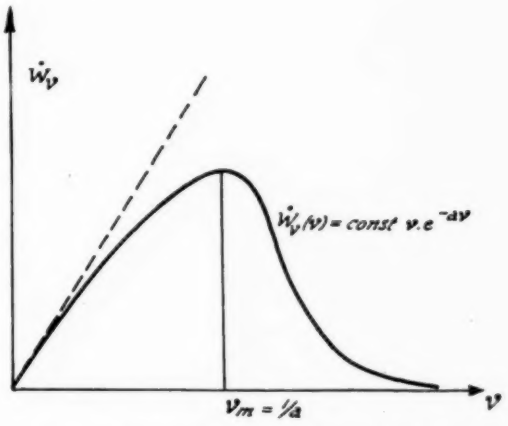


FIG. 8. The functional dependence of W_ν on ν.

In the above investigation we have considered an infinite semispace bound by a plane. If this semispace becomes a finite region the spectrum of its eigenvalues is not continuous, but it is a line spectrum. However, only finite objects are available to us in nature. Therefore, in order to generate a line spectrum, e.g., in the millimeter wave region, one may use a suitable dielectric (or ferromagnetic) resonator. It is not necessary to work with dielectrics exhibiting anomalous dispersion at the appropriate wavelength.

ACKNOWLEDGMENTS

The author wishes to thank Dr. D. Gabor of the Imperial College, London for encouragement and for valuable suggestions.

The author also desires to thank Mr. L. J. Davies, Director of the Research Laboratories of the British Thomson-Houston Company, Ltd., for the permission to publish this work.

APPENDIX I. THE ENERGY TRANSFERRED BY A POINT CHARGE TO AN EIGENFIELD $\bar{A}_\lambda$

The amplitude q_λ of the field $\bar{A}_\lambda$ is a solution of the equation:

$$\ddot{q}_\lambda+\nu_\lambda^2 q_\lambda=\frac{e}{c}\bar{v}\cdot\bar{A}_\lambda,$$

which is an equation of forced oscillations, the applied force being $F_\lambda=(e/c)\bar{v}\cdot\bar{A}_\lambda$. Thus it follows that the work done by the force F_λ during a time interval dt is:

$$dU_\lambda=F_\lambda\cdot\dot{q}_\lambda\cdot dt$$

or the power output

$$dU_\lambda/dt=W_\lambda=F_\lambda\cdot\dot{q}_\lambda$$

which is the required formula.[4]

[4] See also reference 1, p. 55.

Silicon Solar Energy Converters

M. B. PRINCE
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received August 12, 1954)

Theory is given for the design of silicon solar energy converters commonly known as the Bell Solar Battery. Values are given for the various parameters in the design theory. Experimental data are presented and compared with the theoretical relations based on a simple model.

It is found that with present techniques, units can be made with up to 6 percent efficiency in the conversion of solar radiant energy to electrical energy. An important factor in obtaining such high efficiencies is the reduction of the series resistance of the cell to as low a value as possible.

INTRODUCTION

WITH the development of the technique of solid diffusion in producing *p-n* junctions in silicon,[1] new silicon devices[2,3] are now possible for the first time. Among these is a large area silicon photodiode or solar energy converter[3] commonly known as the Bell Solar Battery which has a higher conversion efficiency of solar radiant energy to electrical energy than any other device.

This paper will present the theory for the design of solar energy converters. General relations will be derived from which it will be shown that with the present status of semiconductors, silicon is the best material available from which efficient solar energy converters can be made for use at temperatures near 300°K.

With the use of reasonable anticipated values of various parameters, it will be shown that silicon solar energy converters should be able to be made with efficiencies as high as 10 percent. At present, the best units are 6 percent efficient.[3] Reasons for this result will be given.

GENERAL THEORY

The problem of the photoelectric power converter may best be defined by considering the case of an ideal *p-n* junction with a constant current source in parallel with the junction. The constant current source results from the excitation of an excess over the thermal value of electrons into the conduction band and the diffusion and drift of these carriers across the barrier. The nature of the source need not concern us; it may be a photon source (solar radiation, γ radiation, incandescent lamp, x-rays, etc.) a high-energy particle source (electron gun, β-radioactive elements,[4] α particles, protons, neutrons, etc.), or any other means for creating electron-hole pairs without changing the properties of the ideal junction appreciably (that is, without creating a large electric field or high temperatures in the vicinity of the junction).

The $I-V$ characteristic of such a device is given as

$$I=I_0(e^{qV/kT}-1)-I_L, \tag{1}$$

where I_0 = reverse saturation current of ideal junction; q = charge of electron, k = Boltzmann's constant, T = absolute temperature, and I_L = strength of constant current source.

A plot of Eq. (1) is given in Fig. 1 for selected values of the parameters. It is seen that the curve passes through the fourth quadrant and therefore that power can be extracted from the device. By properly choosing a load, it is possible to extract close to 80 percent of the product $I_{sc}\times V_{oc}$ where I_{sc} is the short circuit current and V_{oc} is the open circuit voltage of the device.

Let us calculate the maximum power that can be obtained from a solar energy converter exclusive of losses by recombination and series and shunt resistances. In bright sunlight at sea level, if every photon falling on a unit created one hole-electron pair that caused

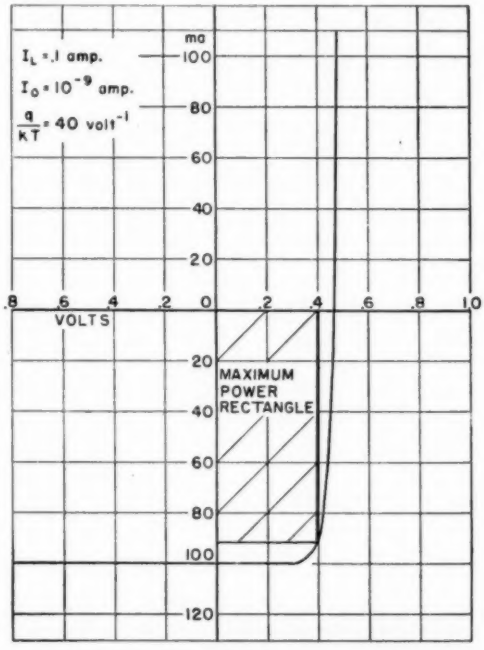


FIG. 1. Ideal $I-V$ characteristic of a photocell or solar energy converter.

[1] C. S. Fuller (private communication).
[2] G. L. Pearson and C. S. Fuller, Proc. Inst. Radio Engrs. **42**, 760 (1954).
[3] Chapin, Fuller, and Pearson, J. Appl. Phys. **25**, 676 (1954).
[4] P. Rappaport, Phys. Rev. **93**, 246 (1954); W. G. Pfann and W. van Roosbroeck, J. Appl. Phys. **25**, 1422 (1954).

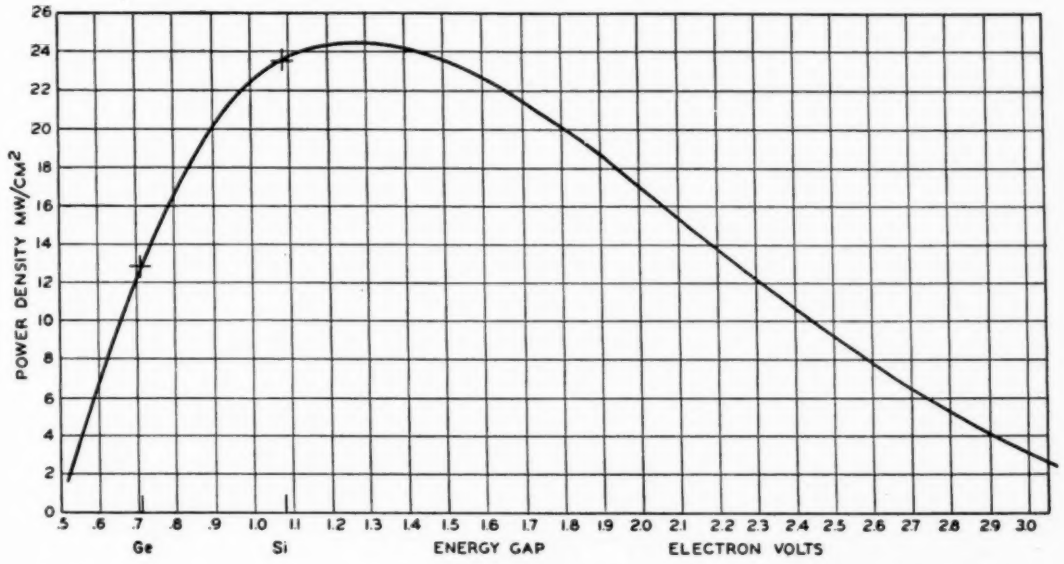


FIG. 2. Maximum converted power density in bright sunlight as a function of energy gap of semiconductor.

current to flow across the junction, there would be a short-circuit current of 0.080 ampere per square centimeter of effective device area. However, the long wavelength limit for the creation of hole-electron pairs by photons depends on the energy gap of the semiconductor involved and this reduces the magnitude of the highest possible short-circuit current to 0.044 ampere/cm² in silicon and 0.068 ampere/cm² in germanium. Further reduction of these values occurs due to reflection losses. Even with "nonreflective" coatings the maximum short-circuit currents will be near 0.035 ampere/cm² in silicon and 0.055 ampere/cm² in germanium. More will be said about this possibility of improving present efficiencies later in the paper. One might think that by choosing semiconductors with lower energy gaps, one would obtain larger short-circuit currents and thus higher efficiencies. However, as will be shown in the next paragraph, decreasing the energy gap reduces the open-circuit voltage at a much faster rate than the short circuit current increases.

The open-circuit voltage V_{oc} at room temperature is given by

$$V_{oc}=0.0575\log_{10}(I_L/I_0)\text{ volts.} \tag{2}$$

Therefore V_{oc} can be maximized by minimizing I_0. I_0 is approximately given by

$$I_0=\left\{qp_n\left(\frac{D_P}{\tau_P}\right)^{\frac{1}{2}}+qn_P\left(\frac{D_n}{\tau_n}\right)^{\frac{1}{2}}\right\}\times\text{area} \tag{3}$$

where p_n=equilibrium density of holes in n-region; n_p=equilibrium density of electrons in p-region; D_p, τ_p=diffusion constant and lifetime of holes in n-region; and D_n, τ_n=diffusion constant and lifetime of electrons in p-region. Consider a heavily doped p-region in contact with a moderately doped n-region, a favorable condition for low I_0 using the diffusion technique. Then

$$J_0=\frac{I_0}{\text{area}}\simeq qp_n\left(\frac{D_p}{\tau_p}\right)^{\frac{1}{2}}, \tag{3a}$$

using the relationships

$$p_n=\frac{n_i^2}{n_n}=\frac{2.23\times10^{31}T^3e^{-E_g/kT}}{n_n} \tag{4}$$

and

$$n_n=\frac{1}{\rho_n q\mu_n} \tag{5}$$

where n_i=equilibrium density of electrons in intrinsic semiconductor, n_n=equilibrium density of electrons in n-region, ρ_n=resistivity of n-region, μ_n=mobility of electrons in n-region, and E_g=energy gap of semiconductor (volts). The open-circuit voltage can be expressed at 300°K as

$$V_{oc}=0.0575\log_{10}J_L\left\{\frac{0.062e^{39E_g}}{\rho_n\mu_n}\left(\frac{\tau_P}{D_p}\right)^{\frac{1}{2}}\right\}, \tag{6}$$

where J_L is the light current density.

It is seen from Eq. (6) that in order to obtain a large V_{oc} one should use a semiconductor with a large energy gap, low resistivity consistent with the diffusion process, material with low mobility, and high lifetime. Assuming a lifetime of ten microseconds, a resistivity of one-tenth ohm centimeters, a mobility variation of $\mu_nE_g^{2.5}$=constant, and a diffusion constant variation of $D_PE_g^{2.5}$=constant, one can calculate the maximum power density as a function of the energy gap. The last two assumptions are not important in that almost any other assumption would give results with less than a few percent difference from the results given in Fig. 2. It is seen that for optimum power with respect to the solar spectrum the energy gap in the semiconductor

TABLE I.

Item	Best value	Limiting factor
$J_{sc}=J_L$	25 ma/cm²	reflection
τ_p	10^{-7} sec	diffusion technique
ρ	0.1 ohm-cm	surface solubility of B in Si is about 10^{18} cm^{-3}
D_p	10 cm²/sec	consistency with high τ_p and ρ[a]
μ_n	700 $\frac{\text{cm}^2}{\text{volt-sec}}$	consistency with high τ_p and ρ[a]

[a] M. B. Prince, Phys. Rev. **93**, 1204 (1954).

should lie between 1.0 and 1.6 electron volts. Thus silicon with a room temperature energy gap of 1.08 electron volts is ideally suitable as a material for solar energy converters. It is interesting to note that similar devices made from germanium would have about half the conversion efficiency of silicon devices. Since the solar radiant power density is 108 milliwatts per square centimeter at sea level on a bright clear day with the sun at the zenith and maximum power density that can be obtained from a silicon unit is 23.5 milliwatts per square centimeter, the maximum possible efficiency for a silicon energy converter would be 21.7 percent.

It is the purpose of this paper to explain why this efficiency is unattainable and to predict maximum designable efficiencies.

THEORY OF SILICON SOLAR ENERGY CONVERTER

As has been mentioned in the last section, the maximum I_{sc} in silicon solar energy converters even with "nonreflective" coatings is about 35 milliamperes/cm². This fact immediately reduces the maximum efficienty to 17.2 percent or 18.6 milliwatts/cm². However these figures are based on anticipated values of various parameters; that is, improvement in the quality of the present silicon. The best values at present using the solid diffusion technique of introducing boron into n-type silicon are given in Table I. Using these values, the maximum expected power density P_A is

$$P_A=0.8J_{sc}V_{oc}=0.8\times0.025\times0.58=0.0116 \text{ watts/cm}^2$$

or the maximum expected efficiency is 10.7 percent.

Even 10 percent efficient units have not been made and reasons for this fact must be given. Up to the present discussion only ideal junctions have been considered. Now let us consider a practical unit. It may have some shunt resistance R_{sh} and certainly has some series resistance R_s due to the body material and the contact to the body. It can be readily shown

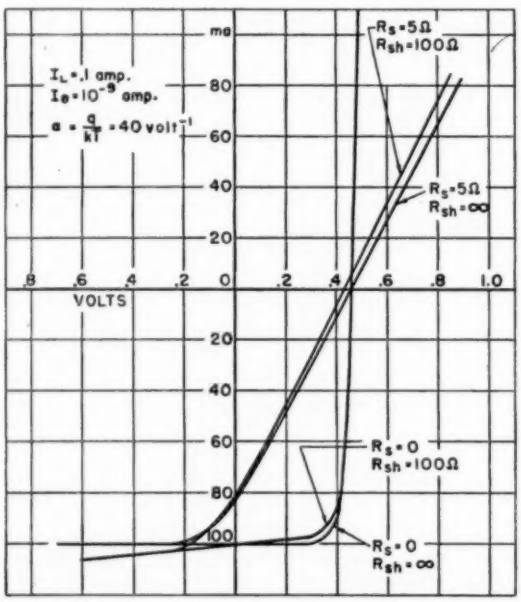


FIG. 3. Theoretical $I-V$ characteristic for various converters that include series and shunt resistances.

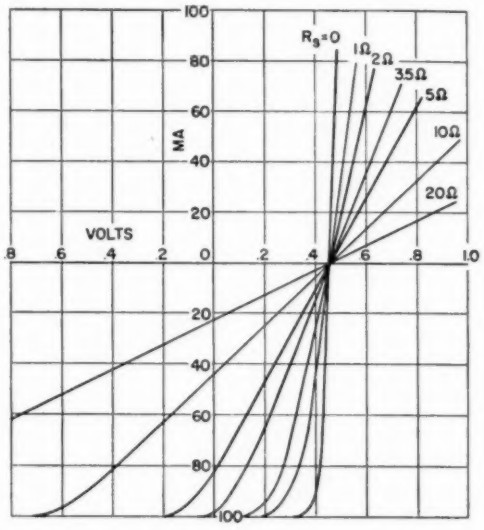


FIG. 4. Theoretical $I-V$ characteristic for various converters with different series resistance.

that for a model containing a R_s and a R_{sh}, the $I-V$ characteristic is given by

$$\ln\left[\frac{I+I_L}{I_0}-\frac{V-IR_s}{I_0R_{sh}}+1\right]=\frac{q}{kT}(V-IR_s). \qquad (7)$$

Plots of this equation with all combinations of $R_s=0$, 5 and $R_{sh}=\infty$, 100 ohms are given in Fig. 3 with the same parameters I_0, I_L, and kT/q as in Fig. 1. It can be seen that a shunt resistance even as low as 100 ohms does not appreciably change the power output of a unit whereas a series resistance of only 5 ohms reduces the available power to less than 30 percent of the optimum power with $R_s=0$. Since the R_{sh} does not affect our results, let us neglect it in future calculations. Figure 4 shows the theoretical $I-V$ characteristics of units having $R_s=0$, 1, 2, 3.5, 5, 10, and 20 ohms. Table II gives the ratio of the relative maximum available power from such units. These data are plotted in Fig. 5 which shows graphically how extremely

important it is to reduce the series resistance to as low a value as possible.

Now that it has been shown that the series resistance is a controlling parameter in reducing the available power from a silicon solar energy converter, let us consider a possible design of a unit and place some limits on the controlling parameters.

Consider Fig. 6 which shows a possible configuration for a solar energy converter. Radiation is incident on the top surface. Contacts are made to the n island and p ring on the bottom surface. The geometrical parameters are the length L, width W, thickness H, and the depth of the p-type layer t. The length and width of the unit are limited by the size of crystal from which the unit is cut. At the present time L is limited to about 6 cm and W to 1.5 cm. The thickness of the wafer H should be made small such that the resistance in the n-type region of the device is a minimum. However this parameter is not critical since it is easy

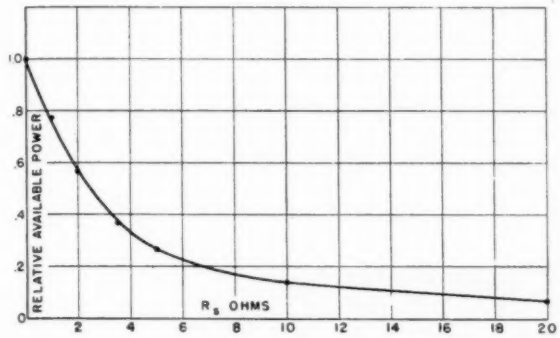


FIG. 5. Relative maximum available power as a function of the series resistance.

to keep the resistance of this part of device below $\frac{1}{4}$ ohm with thicknesses between 0.1 and 0.5 cm. Therefore in the interest of conservation of material, it is desirable to make $H \simeq 0.1$ cm. A more critical dimension is the thickness of the p-layer t. Practically all of the electron-hole pairs produced by sunlight are within 10^{-4} cm of the surface. Thus two conflicting demands are made on t; thick t to reduce the resistance of the layer and thin t to prevent loss of minority carriers by recombination before diffusing to the junction. It can be shown that the p-layer resistance is given by

$$R = W\rho/4tL. \tag{8}$$

The power that can be obtained from such a device is given by

$$P = -IV = -I\left[\frac{kT}{q}\ln\left(\frac{I+I_L}{I_0}+1\right)+IR_s\right]. \tag{9}$$

For the purposes of calculations let us assume the following values of some of the parameters:

$p_n = 10^5$ cm^{-3} $\quad kT/q = 0.025$ volt

$D_p = 10$ cm^2/sec

$\tau_p = 10^{-7}$ sec

TABLE II.

R_s	Relative maximum available power
0 ohms	1
1	0.77
2	0.57
3.5	0.37
5	0.27
10	0.14
20	0.07

thus

$J_0 = 10^{-10}$ amp/cm^2 = reverse saturation current density,

$J_{L0} = 0.03$ amp/cm^2 = light current density at surface of unit,

$L_D = 10^{-3}$ cm = diffusion length in p-layer.

Therefore $I_0 = J_0 WL = 10^{-10}WL$ amp and $I_L \cong J_{L0} \times WLe^{-t/L_D} = 0.03WL\exp(-10^3 t)$ amp. The last relation is a good approximation for $t \geq L_D$. For $t \ll L_D$, the approximation fails but in the following calculations, the error introduced by the approximation is negligible. Substituting these values and (8) into (9), we find

$$P = -0.025I\ln\left[\frac{I+0.03WL\exp(-10^3 t)}{10^{-10}WL}+1\right] - I^2\frac{W}{4}\frac{\rho}{tL}. \tag{10}$$

The first term represents the actual power developed and the second term represents the power lost in the p-type layer. Equation (10) indicates that one should make L as large as possible and W small. Let $L = 5$ cm as this is a convenient limit for this parameter and $W = 1$ cm as anything smaller will make the fabrication unnecessarily difficult. Since the surface resistivity of

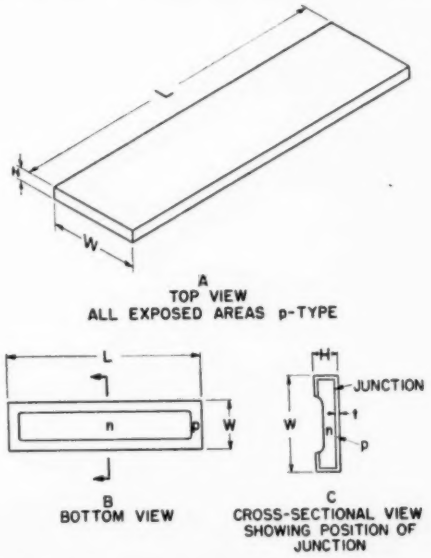


FIG. 6. Geometry of solar energy converter.

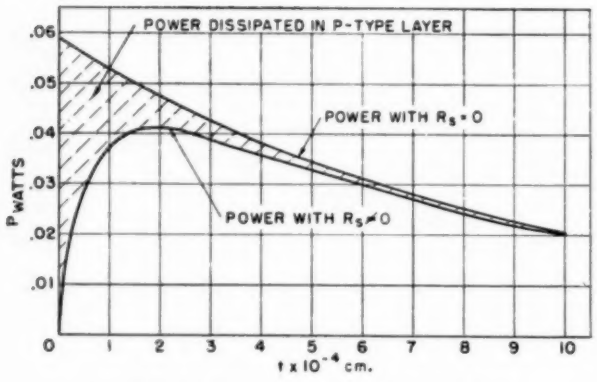


FIG. 7. Maximum available power as a function of active p-layer thickness.

boron diffused silicon is about 10^{-3} ohm-cm,[1] let ρ be 2×10^{-3} ohm-cm as an average resistivity of the p-type layer. Then (10) can be rewritten as

$$P=-0.0575I\log[2\times10^{9}I+3\times10^{8}\exp(-10^{3}t)]-10^{-4}\frac{I^{2}}{t}. \quad (11)$$

Equation (11) has been maximized with respect to I for various values of t and plotted in Fig. 7. The upper curve gives the maximum value of the power for the first term in the above expression and the lower curve represents the maximum value for the entire expression. It is observed that the maximum power occurs for our chosen geometry with $t=2\times10^{-4}$ cm and that the maximum is quite broad. For this unit (5 cm^2) there is an efficiency of over 8 percent and a p-layer resistance of 0.5 ohm.

On a unit area basis, Eq. (10) can be written as

$$P_A=\frac{P}{WL}=-0.025J\times\ln\left[\frac{J+0.03\exp(-10^{3}t)}{10^{-10}}+1\right]-J^{2}\frac{W^{2}\rho}{4t}, \quad (12)$$

$$=-0.0575J\log[10^{10}J+3\times10^{8}\exp(-10^{3}t)]-J^{2}\frac{W^{2}\rho}{4t}. \quad (12a)$$

Equation (12) indicates that the power per unit area developed is independent of the length of the specimen but depends on the square of the width. A plot of maximum power per unit area *versus* the width of specimen is given in Fig. 8 along with a plot of the product as a function of the width. It is seen in Fig. 8 that with our assumptions and from any standpoint, one would not want to make any units with widths greater than about 1.7 cm. If one wants a unit with high photosensitivity (high efficiency), the width should be made as small as possible whereas if one wants as much power as possible from a single unit then the width should be made about 1.5 cm. A more accurate calculation indicates that the product of $P_A\times W$ approaches an asymptotic value and does not fall off with W as is shown in Fig. 8.

Another serious source of series resistance is in the actual contact of the conductors to the n- and p-type regions of the silicon.

The temperature dependence of the operating characteristics is always of interest in the description of semiconductor devices. Since the short circuit current depends only on the light current intensity, it will have no temperature dependence (except for minor corrections due to lifetime changes and series resistance changes). The most important temperature effect is through the change in open circuit voltage. It can be shown that the temperature variation of the open circuit voltage can be represented in silicon by

$$dV_{oc}/dT=-0.00288 \text{ volt/°C}. \quad (13)$$

Since the output power of the device varies linearly with the open-circuit voltage, the power decreases with temperature at a rate greater than ½ percent per degree centigrade. It will be shown in the next section that Eq. (13) is satisfied by experiment.

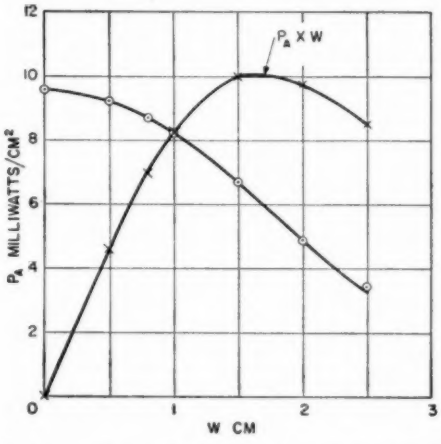


FIG. 8. Maximum available power density as a function of width of solar energy converter.

RESULTS OF EXPERIMENTAL MEASUREMENTS

The results to be presented in this section consist of the following:

1. The $I-V$ characteristic of a unit before and after reducing the series resistance of the device. 2. The $I-V$ characteristic of a unit as a function of the light current. 3. The $I-V$ characteristic of 10 units connected in series. 4. The open circuit voltage as a function of temperature.

The units whose properties are given below were made by diffusing boron into n-type silicon. The dimensions of these units are L=5.7 cm, W=1.2 cm, H=0.1 cm, and t=0.0002 cm.

Figure 9 shows the results of decreasing the series resistance of a cell on the $I-V$ characteristic of the

converter. The sample originally had a series resistance of 6.1 ohms. After the contact was improved the series resistance dropped to about 2.7 ohms. This decrease in series resistance allowed one to obtain from the unit at optimum load 2.2 times as much power after the improvement compared to that obtained before the improvement. Further reduction of the series resistance of this converter would lead to further increase in the maximum available power. This unit has a shunt resistance of the order of 1000 ohms.

The $I-V$ characteristics of a converter as a function of the light intensity are given in Fig. 10. This unit has a series resistance of 1.8 ohms and on comparing the 102 ma short-circuit current curve with those of Fig. 4, it is seen that the experimental curve fits the

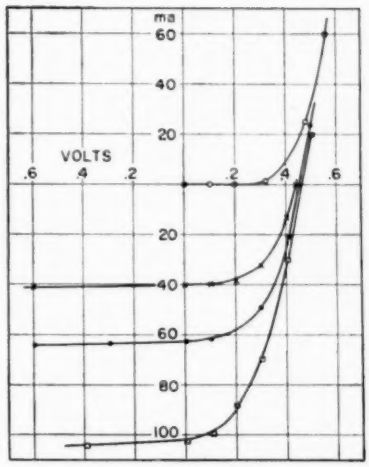


FIG. 10. Family of experimental $I-V$ characteristics of a solar energy converter for various light intensities.

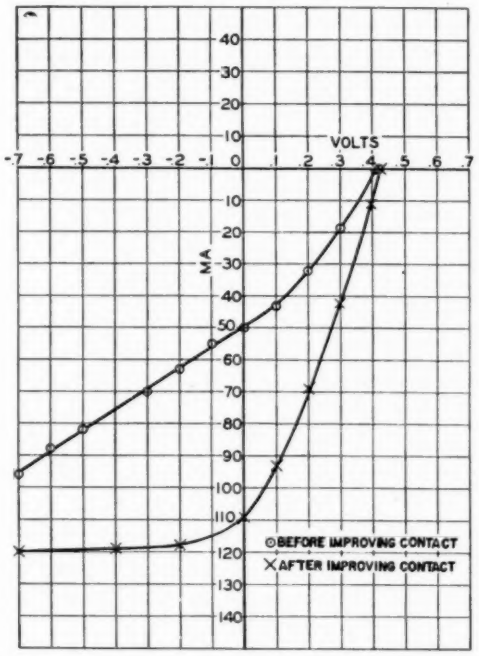


FIG. 9. Experimental $I-V$ characteristics of a converter before and after improving contact.

theoretical curve extremely well. It has been observed on this and on other units that the maximum available power under different intensities of illumination occurs at a constant voltage for a particular converter. In the case of the unit whose characteristics are illustrated in Fig. 10, this voltage is 0.30 volt or $\frac{2}{3}$ the open circuit voltage. As a consequence of this fact, the ideal load for a solar converter would be one that required a constant voltage; e.g., a storage battery. This unit also has a shunt resistance of 1000 ohms

The $I-V$ characteristic for a series connection of 10 solar converters in bright sunlight is given in Fig. 11. The series resistance of this collection is 18.3 ohms or an average of less than two ohms per cell. The maximum power is obtained at a voltage equal to $\frac{2}{3}$ the open-circuit voltage and is equal to 0.2 watt. Since each unit has an area of 6.8–7.0 cm^2, the device is only about 3 percent efficient in the conversion of sunlight to electrical power. Individual units have been made with 5 percent to 6 percent conversion efficiencies.[3]

Figure 12 gives an experimental plot of the open-circuit voltage as a function of temperature. The straight line through the data has a slope of -0.00288 volt per degree centigrade in agreement with the value given by Eq. (13). The disagreement of the low-temperature point is probably due to the fact that the sample was cooled with the aid of dry ice and the temperature read by thermocouple system was not the effective average temperature of the device.

CONCLUSIONS

The experimental results given in the last section are in agreement with the design theory for boron diffused silicon solar converters. This fact leads us to believe that one should be able to produce units for conversion of incident solar radiant energy into electrical energy with efficiencies as high as 10 percent with modifications of present techniques. Cells have been made that have conversion efficiencies up to 6 percent.

Theoretical predictions and experimental confirmations indicate that the most important factor in the design of a solar energy converter is the series resistance

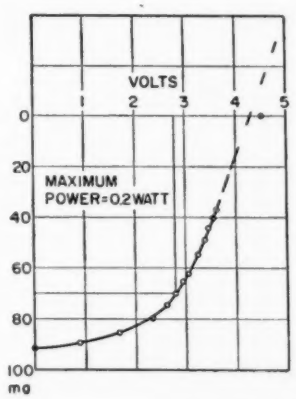


FIG. 11. $I-V$ characteristic of a series connection of ten solar energy converters in bright sunlight.

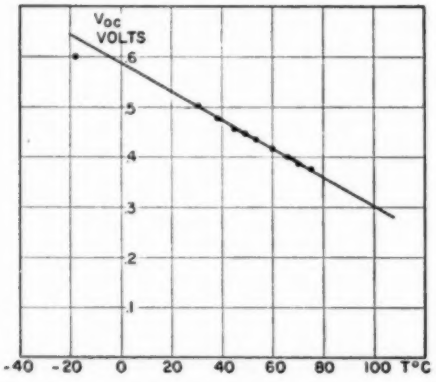


FIG. 12. Open-circuit voltage as a function of temperature.

of the device. Experimentally, units have been made with the series resistance being less than two ohms. Design theory indicates that units should be able to be made having a series resistance of less than one ohm. An experimental attainment of this value would improve the output of these units by 30 to 40 percent or to about 7 or 8 percent efficient.

Another method for possibly increasing the output of these devices appreciably would be the use of nonreflective coatings either of the $\frac{1}{4}\lambda$ type or by the use of various materials having intermediate indices of refraction between silicon ($n=4$) and air ($n=1$). With this thought in mind a unit was cemented to soft glass ($n=1.4$) with polystyrene cement ($n\cong 2$). The unit did not show any improvement. There are two possible explanations for this behavior. First, there is always a thin layer of quartz ($n=2$) on the silicon which makes the improvement by the addition of other films of negligible importance. The second possibility is that any improvement in the reflection loss is balanced by an absorbtion loss in the glass and cement. Probably the experimental facts are the result of a combination of these explanations.

Other possible means for increasing the output of these solar energy converters is by the use of reflectors (mirrors) or concentrators (lenses) whereby the intensity of radiation is increased. However these methods would be practical only in the case of small power supplies.

The author wishes to acknowledge several stimulating discussions with J. L. Moll, G. L. Pearson, D. M. Chapin, and C. S. Fuller. Much of the fabrication of the devices was done by T. J. Vasko.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Effect of Air Damping on Transverse Vibrations of Stretched Filaments*

D. W. STAUFF AND D. J. MONTGOMERY
Michigan State College, East Lansing, Michigan

(Received September 20, 1954)

The effect of air damping on the resonant frequencies of forced vibration of stretched filaments is studied by observing the shift in frequency when the surrounding medium is changed from air at atmospheric pressure to hydrogen at the same pressure, or to air at reduced pressures. For fibers of linear density less than about 10 micrograms per centimeter, the shift between atmospheric air and vacuum exceeds one cycle per second, becoming greater as the fibers become finer. The frequency shifts observed experimentally are compared with those predicted by certain formulas of Stokes. It is found that the theory works well enough to allow the effect of air damping to be taken into account in measurements of linear density by means of the vibroscope. However, the observed shift seems to be about one or two cycles per second higher than that calculated in the case of small shifts, and about 20 or 30 percent higher in the case of large shifts. A rigorous evaluation of the applicability of Stokes' theory would require further investigation.

INTRODUCTION

IN the vibroscopic method of determining the mean cross-sectional area of filaments,[1–5] the elementary formulas of Mersenne and Galileo must be modified to take into account stiffness and nonuniformity. Studies of these effects are available.[6,7] With very fine filaments another effect, that of damping by the surrounding medium, becomes important. To study this effect experimentally, filaments of various linear densities were vibrated in air and in hydrogen at atmospheric pressure, and in air at reduced pressure, the shift in resonant frequency upon change of medium being observed. To get help from theory, the analysis of Stokes[8] for a related case has been applied.

* A portion of a thesis submitted by the first-named author to the School of Graduate Studies of Michigan State College in partial fulfillment of the requirements for the M.S. degree.

[1] V. E. Gonsalves, Textile Research J. **17**, 369 (1947).
[2] S. L. Dart and L. E. Peterson, Textile Research J. **19**, 89 (1949); **22**, 819 (1952).
[3] D. O. Sproule, J. Textile Inst. **43**, T455 (1952).
[4] D. J. Montgomery and W. T. Milloway, Textile Research J. **22**, 729 (1952).
[5] J. Lindberg, Textile Research J. **23**, 67 (1953).
[6] D. J. Montgomery, J. Appl. Phys. **24**, 1092 (1953).
[7] E. T. L. Voong and D. J. Montgomery, Textile Research J. **23**, 821 (1953).
[8] (a) G. C. Stokes, Trans. Cambridge Phil. Soc. **9**, 8–106 (1856); (b) *Mathematical and Physical Papers* (Cambridge

THEORY

According to Stokes, the damping force per unit length of an infinite rigid cylinder in transverse sinusoidal oscillation in a viscous fluid is given by

$$-\phi M\gamma S\ddot{y}-\alpha\phi M\gamma S\omega\dot{y}, \tag{1}$$

where $\ddot{y}$ and $\dot{y}$ are, respectively, the transverse acceleration and velocity. The dimensionless parameter M is defined by

$$M\equiv 4\mu/a^2\gamma\omega, \tag{2}$$

where μ = viscosity of medium (poise), a = radius of cylinder (cm), γ = density of medium (g/cm³), and $\omega=2\pi f$ = angular frequency of transverse displacement (rad/sec). Further, $S=\pi a^2$ = cross-sectional area of cylinder (cm²).

The quantities α and ϕ are complicated dimensionless functions† of M, given by Stokes for small values of M in series form, and for large values ($M>100$) in the approximate form

$$\alpha=(4/\pi)(\tfrac{1}{2}\ln M-0.577), \tag{3}$$

and

$$\phi=\frac{4\pi}{16(0.577-\frac{1}{2}\ln M)^2+\pi^2}. \tag{4}$$

These functions are plotted in Fig. 1.

From Eq. (1) it is seen that the effect of the surrounding medium shows up under two aspects: a viscous damping (i.e., proportional to dy/dt); and an inertial drag (i.e., proportional to d^2y/dt^2). Kärrholm and Schröder,[9] following Ising,[10] neglected the inertial term, basing their procedure on a remark of Stokes[11]: ". . . it would seem that when the radius of the cylinder is very small, the motion which would be expressed by the formulae [of Eq. (1)] would be unstable. If so, the quantity of fluid carried by the wire would be diminished, portions being continually left behind and forming eddies. The resistance which would be a function of the velocity . . ." However, in our experiments the Reynolds-Stokes number is very small and consequently, we doubt if eddies are formed; moreover, the results of the calculations upon neglect of the inertial term give predicted values much too low. Hence, we believe that the inertial term must be included. On the other hand, from our results we cannot rule out the possibility that the damping force consists only of a term proportional to the first derivative, but with a different factor.

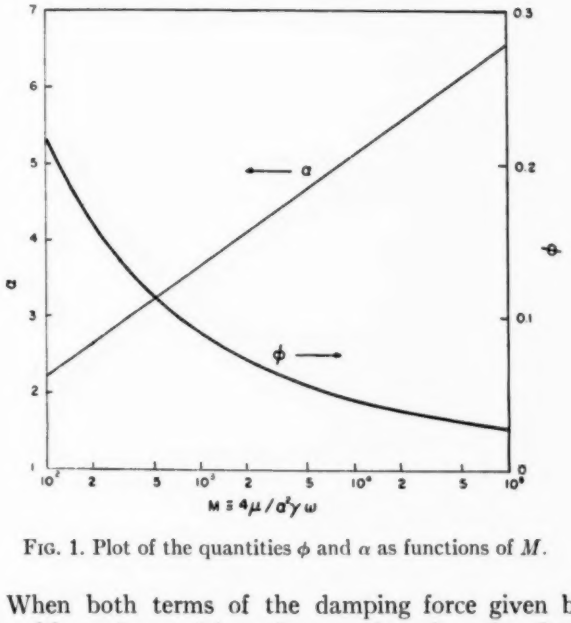


FIG. 1. Plot of the quantities ϕ and α as functions of M.

When both terms of the damping force given by Eq. (1) are inserted into the equation of motion for a vibrating filament, the result is the following‡:

$$Ty''-\alpha\phi M\gamma S\omega\dot{y}+F(x)e^{i\omega t}=(\rho S+\phi M\gamma S)\ddot{y}, \tag{5}$$

where $y=y(x,t)$ = transverse displacement (cm), $\ddot{y}=\partial^2y/\partial t^2$, where t is the time (sec), $y''=\partial^2y/\partial x^2$, where x is the distance along the filament (cm), T = tension in filament (dynes), ρ = density of filament (g/cm³), and $F(x)e^{i\omega t}$ = driving force per unit length (dyne/cm).

The solution of Eq. (5) for the displacement y is

$$y(x,t)=\frac{e^{i\omega t}\sum F_n e^{in\pi x/l}}{T(n\pi/l)^2+[i\alpha\phi M\gamma S\omega^2-\omega^2(\rho+\phi M\gamma)S]}, \tag{6}$$

where l is the length of the filament (cm) and F_n is the nth Fourier space component of the driving force.

The amplitude of the curve $y(x,t)$ at given x will have its maximum, as a function of ω, when the absolute value of the denominator in Eq. (6) is a minimum. The equation resulting from setting the derivative of the absolute value of this denominator to zero is intractable, but if the slowly varying functions ϕ and α are treated as constants, the resulting equation may be easily solved and the resonant frequencies therby obtained to good approximation. If we introduce ω_0, the fundamental resonant frequency in the absence of damping, by the definition

$$\omega_0^2\equiv(\pi/l)^2(T/\rho S), \tag{7}$$

and define ϵ as the fractional deviation of $\omega_n\equiv 2\pi f_n$, the nth resonant frequency in the presence of damping, as follows:

$$\omega_n\equiv 2\pi f_n=n\omega_0(1-\epsilon), \tag{8}$$

University Press, Cambridge, England, 1901), reprint, Vol. III, pp. 1–141.

† For our purposes it has been convenient to replace the k and k' of Stokes' work by the related functions α and ϕ. Moreover, we have substituted M for his $1/m^2$.

[9] E. M. Kärrholm and B. Schröder, Textile Research J. **23**, 208 (1953).

[10] G. Ising, *Undersökning Rörande Elektrometrar* (Glecrupska Universitets Bokhandeln, Lund, Sweden, 1919), part II, pp. 281–3. We are grateful to Mrs. Kärrholm and Mr. Schröder for supplying us with a photostat of the relevant portion of the work.

[11] Reference 8(a), p. 57, or 8(b), p. 67.

‡ The internal (frictional) damping in the fiber is neglected in comparison with the external damping; this assumption is justified when vibrating under atmospheric pressure.

then to first-order terms in ϵ it is found that

$$\epsilon=(\phi\mu/f_n\rho S)^2(\alpha^2+1)+(\phi\mu/f_n\rho S). \quad (9)$$

It turns out that the two terms in Eq. (9) are of the same order of magnitude, and hence, an upper limit on the frequency shift Δf_n is

$$\Delta f_n \equiv \epsilon f_n \sim 2\phi\mu/\rho S. \quad (10)$$

For vibroscoping in air under ordinary conditions φ lies around 0.1, and hence for the condition $\Delta f_n=1$ cycle per second, the linear density ρS must be about 10–20 micrograms per centimeter. Thus, to get a readily measurable shift, it is necessary to work with filaments finer than 10 μg/cm, corresponding to diameters of less than 0.0010 in. for glass or aluminum, 0.0005 in. for copper, and 0.0003 in. for gold or tungsten.

APPARATUS AND PROCEDURE

The electrostatic form of vibroscope is essentially a string electrometer, as described by Montgomery and Milloway.[4] In the present experiments the fiber was suspended inside a glass tube to permit observations in different gases at various pressures. The fiber was observed by means of a cathetometer telescope. A plenum chamber and capillary were included in the system to isolate the fiber from the pressure fluctuations set up by the mechanical vacuum pump, and columns of drierite ($CaSO_4$) and anhydrone ($MgClO_4$) were incorporated to insure the dryness of the gas admitted. Three pairs of curved brass electrodes, diametrically opposed on the outside of the tube, were connected to a step-up transformer driven by a variable-frequency audio oscillator through a circuit described in reference 4. A switchboard made it possible to apply the output of the oscillator to any pair of plates in space and time relation suitable for exciting specific modes of vibration. The filaments used were of tungsten (diam 10μ; density 19.0 g/cm^3), glass (diam 10μ; density 3.2 g/cm^3), copper (diam 10μ; density 8.9 g/cm^3), and gold (diam 8μ, 5μ, 2μ; density 19.3 g/cm^3). It was attempted to work with aluminum, but this element cannot be bare drawn to the necessary fineness, and we were unsuccessful in removing the copper jacket from the particular sample of Wollaston wire available.

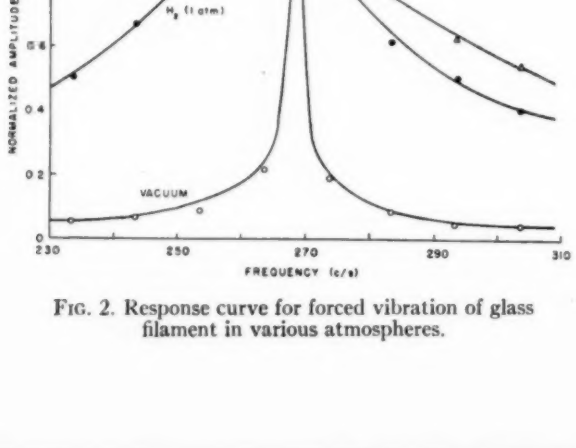


FIG. 2. Response curve for forced vibration of glass filament in various atmospheres.

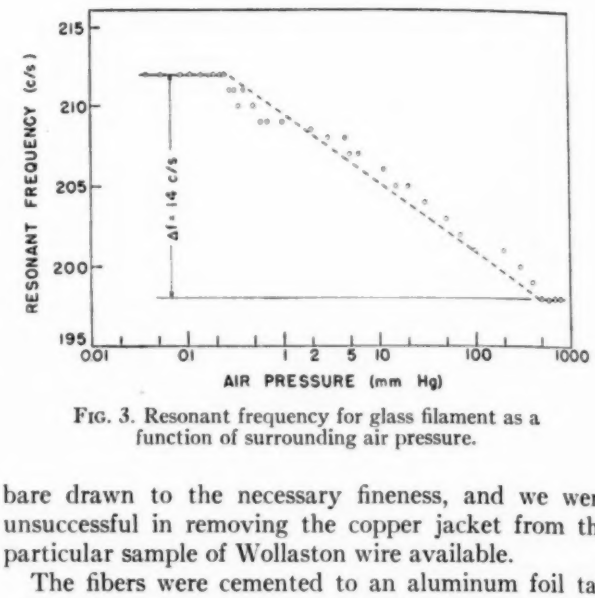


FIG. 3. Resonant frequency for glass filament as a function of surrounding air pressure.

The fibers were cemented to an aluminum foil tab at the bottom and clamped in the brass clamp at the top. Tension was applied by the weight of the tab, augmented when desired by hanging copper-wire hooks through a hole in the tab. An analytical balance was used to weight the tabs and hooks.

Tungsten filaments were best observed when placed directly in front of a microscope lamp as a source of illumination. Glass filaments were seen most readily by refracted light, and the illumination was therefore directed at the fiber obliquely to the line of sight. Gold and copper showed best under reflected light, the source here being a small desk lamp beside the cathetometer. The length of the fiber was measured with the cathetometer, and the amplitude of vibration was measured on a reticle of the cathetometer telescope.

Amplitudes of vibration were purposely kept small, in conformity with the assumptions of the theory. It was found that a slight change in the resonant frequency indeed occurred if the amplitude became very large, and accordingly the transverse displacement of the fiber was kept less than 0.1 mm from equilibrium position, at the mid-point distance of five centimeters from the end of the fiber.

RESULTS

Figure 2 is an example of the observations. Normalized amplitudes of vibration of a glass filament vibrated in air at one atmosphere, in hydrogen at one atmosphere, and in vacuum (29 μHg), are plotted as a function of frequency. A shift to higher frequency and a sharpening of the resonant peak are noted when the viscosity of the surrounding fluid is decreased. The shift is quite well defined in the case of glass, the resonant frequencies being 252, 263, and 268 cps for air, hydrogen, and vacuum, respectively. The normalization of the amplitude to unity at resonance conceals the great increase in amplitude in vacuum over that in air at normal pressure.

Since the viscosity of a gas is nearly independent of density so long as the mean free path is small compared with the relevant dimension of a body in motion with respect to the gas, it is necessary to go to quite low pressures to insure that the shift is complete. To see in detail how the shift progresses as the pressure is lowered, Fig. 3 shows the resonant frequency against air pressure for a certain glass fiber. At high pressures, where the mean free path ($\sim 0.1\mu$ at 1 atmos) is small compared with the fiber diameter ($\sim 10\mu$), the viscosity is constant; at low pressure where the mean free path ($\sim 1000\mu$ at 50 μHg) is large compared with the fiber diameter, the viscosity effectively vanishes. Hence the curve would be expected to consist of two flat regions connected by a transition region, and this expectation is confirmed. We conclude then that at pressures less than, say, 100 μHg the viscosity effects vanish.

To make a quantitative check of the theory we use the data given in Fig. 2. The value calculated from Eq. (9) for the shift from air at normal pressure to vacuum is 11 cps, whereas the values observed in Fig. 2 is 16 cps. For the shift from air to hydrogen the calculated value is 9 cps, and the observed value 11 cps. It appears that the theory gives roughly the value of the shift, but that the predicted value is somewhat low.

Table I gives values of calculated and observed frequency shifts for various fibers vibroscoped in air at atmospheric pressure and in vacuum. Table II gives similar data for fibers vibroscoped in hydrogen at atmospheric pressure and in vacuum. Observed shifts are of relatively low precision, even though points on the frequency curves were repeatedly observed to be reproducible to about one cps (except with very fine fibers), because the shift is small in most cases, and is

TABLE I. Resonant frequency shifts from air (1 atmos) to vacuum.

Nominal diameter (in.)	T (mg)	f (cps)	l (cm)	a (μ)	ϵ (%)	Δf calc (cps)	Δf obs (cps)
Gold (19.3 g/cm³)							
0.0003	72.9	99	10.53	5.18	0.7	0.7	2
0.0003	25.6	58	10.53	5.24	1.1	0.7	1
0.0003	25.6	62.5	9.80	5.24	0.9	0.6	1
0.0002	74.1	261	8.00	2.61	1.0	2.6	7
0.0002	25.8	119	10.00	2.72	2.0	2.4	3
0.0001	6.4	239	5.88	1.14	6.9	16.5	20
Tungsten (19.0 g/cm³)							
0.0005	36.6	56	10.13	6.77	0.8	0.4	0
Copper (8.9 g/cm³)							
0.0004	238.4	227	9.45	6.73	0.6	1.3	2
0.0004	147.5	179	9.45	6.70	0.7	1.3	2
0.0004	50.9	106	9.45	6.65	1.1	1.1	2
Glass (2.5 g/cm³)							
0.0004	51.0	268	10.11	4.62	4.2	11.1	16

TABLE II. Resonant frequency shifts from hydrogen (1 atmos) to vacuum.

Nominal diameter (in.)	T (mg)	f (cps)	l (cm)	a (μ)	ϵ (%)	Δf calc (cps)	Δf obs (cps)
Gold (19.3 g/cm³)							
0.0003	25.6	61.5	9.80	5.33	0.4	0.2	1
Tungsten (19.0 g/cm³)							
0.0005	36.6	56	10.13	6.77	0.8	0.4	0
Glass (2.5 g/cm³)							
0.0004	51.0	263	10.11	4.62	0.9	2.4	5

determined as a difference between two large quantities. The oscillator frequency, which was calibrated by observation of Lissajous figures of output against line frequency on a cathode-ray oscilloscope, did not wander much during the course of an experiment. The line frequency, checked against a tuned-circuit frequency meter, did not drift measurably. The diameter of each fiber (column a, Tables I and II) was determined by means of the vibroscope formula, and it was found that in most cases the calculated diameter varied somewhat from the manufacturer's specifications. Length measurements with the cathetometer were probably accurate to 0.5 percent. The cathetometer was capable of much higher precision, but the point of attachment of the fiber to the tab cannot be precisely determined. Errors in the values of density and applied tension were probably not more than 1.0 percent and 0.3 percent, respectively.

Part of the discrepancy may perhaps be attributed to inaccuracy in frequency readings; but the observed values are consistently high, and we must conclude that there exists some effect not considered in the theory. It is unlikely, however, that the error in predicted values will be more than one or two cps low in the case of small shifts, or more than 20 or 30 percent low in the case of large shifts. In actual use of the vibroscope to determine fiber linear densities, we recommend that formula (9) be used to select experimental conditions which render the effects of damping negligible. In practice such conditions can almost always be attained by increasing the tension, thereby increasing f_n without changing Δf_n, and consequently decreasing ϵ to a negligible value.

CONCLUSIONS

Experimental observations of the effect of air damping of transverse vibrations of stretched filaments of glass, tungsten, copper, and gold by the vibroscope method lead to the following conclusions:

1. Air damping will have an effect on the natural frequency and amplitude of vibration of fine fibers. Under ordinary conditions the frequency shift is less

than one cps for fibers of linear density greater than 10 micrograms/cm, but for finer fibers it becomes of increasing importance.

2. Stokes' theory works well enough to allow rough calculation of the effect of air damping in measurements of the resonant frequency made with a vibroscope; but the observed shift seems to be about one or two cps higher than calculated in the case of small shifts or about 20 or 30 percent higher in the case of large shifts.

3. A rigorous evaluation of the applicability of Stokes' theory will require further investigation.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Plasma Ion Oscillations in Electron Beams

K. G. HERNQVIST
RCA Laboratories, Princeton, New Jersey
(Received July 8, 1954)

The frequency of the oscillations observed in ion-neutralized electron streams such as that of an ion-trapping gun or of a tetrode has been shown to agree well with the Langmuir-Tonks theory for plasma ion oscillations. The existence of a dipole mode of oscillation in addition to the longitudinal mode has been observed for a plasma column. A mechanism is proposed for the excitation of these self-sustained oscillations by which the secondary electron current from electrodes bounding the plasma is modulated by the vibration of the ions in such a way as to deliver energy to the ac field associated with the ion oscillations. Experimental observations verify this picture as far as the source of energy is concerned.

I. INTRODUCTION

UNDAMPED oscillations in the frequency region from one-half to a few megacycles have been found to occur both in gas-discharges[1] and in ion-neutralized electron streams[2] and have been shown to be due to ion plasma vibrations. Pierce[3] has treated such a case theoretically and also suggested that some feedback mechanism by means of secondary electrons from electrodes bounding the plasma might lead to self-oscillations. The importance of the boundary conditions of the plasma for the maintenance of the oscillations has been pointed out by Sluzkin and Maydanov[4] in a proposed explanation for electron plasma oscillations, but no detailed treatment of the mechanism by which energy is transferred to the ac field of the plasma is made.

This report describes a study of plasma ion oscillations in ion-neutralized electron beams (hereafter called space-charge-neutralized electron beams or simply SCN-beams). However, the basic ideas and results will, in general, also be applicable to low-pressure gas discharge plasmas where the mean-free path of the electrons is of the same order as, or larger than the inter-electrode spacings. MKS units are used throughout, except when otherwise specified.

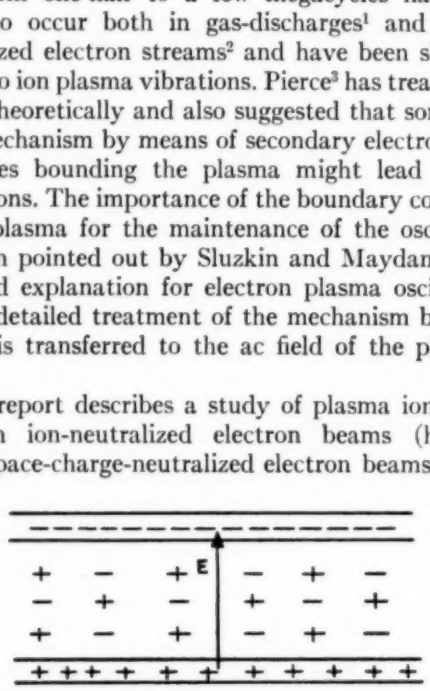


FIG. 1. Oscillations of an infinitely wide plasma disk ($g=1$).

II. THE LANGMUIR-TONKS THEORY FOR PLASMA OSCILLATIONS

An expression for the characteristic frequency of oscillation of a plasma has been derived by Langmuir and Tonks.[5-7] The frequency of ion oscillations is

$$f=\frac{1}{2\pi}g\left(\frac{e^2N}{\epsilon_0(m_i+3\mathcal{C})}\right)^{\frac{1}{2}} \tag{1}$$

where g is a geometry factor, e is the ion charge, N is the charge density, ϵ_0 is the permittivity of free space, m_i is the ion mass, and $3\mathcal{C}$ is a correction factor which is inversely proportional to the electron temperature. For high electron temperatures or for the case of ion vibrations in SCN-beams, $3\mathcal{C}$ is negligible.

Oscillations of an infinitely wide plasma disk (Fig. 1) give rise to a uniform electric field inside the disk, and for this case the geometry factor is $g=1$. Transverse oscillations of a cylindrical plasma column (Fig. 2) give rise to a uniform electric field inside the plasma and a dipole field outside. In this case $g=1/\sqrt{2}$. Similarly for a spherical plasma, $g=1/\sqrt{3}$. Other modes than the principal dipole mode of oscillation are also possible, as has been shown by Herlofson.[8]

[1] N. R. Labrum and E. K. Bigg, Proc. Phys. Soc. (London) **B65**, 365 (1952).
[2] E. G. Linder and K. G. Hernqvist, J. Appl. Phys. **21**, 1088 (1950).
[3] J. R. Pierce, J. Appl. Phys. **19**, 231 (1948).
[4] A. A. Sluzkin and A. P. Maydanov, J. Phys. U. S. S. R. **6**, 7 (1942).
[5] L. Tonks and I. Langmuir, Phys. Rev. **33**, 195 (1929).
[6] L. Tonks, Phys. Rev. **37**, 1458 (1931).
[7] L. Tonks, Phys. Rev. **38**, 1219 (1931).
[8] N. Herlofson, Arkiv Fysik **3**, 247 (1951).

For SCN-beams the ion density N is approximately equal to the electron density of the electron beam and Eq. (1) can be written

$$f=0.679\times10^6 g\left(\frac{j}{M\sqrt{V}}\right)^{\frac{1}{2}} \tag{2}$$

where j is the current density, V is the electron-beam voltage, and M is the molecular weight of the ions.

III. PLASMA BOUNDARY CONDITIONS FOR THE SCN-BEAM

It has been shown[2,9] that if an electron stream is injected into a region which has no electric fields transverse to the electron stream and which is bounded by electrodes at potentials well above the ionization potential of the residual gas, ions will be trapped in the electron space-charge field and a special kind of plasma is formed. This plasma differs from an ordinary gas-discharge plasma in that the electrons travel through the ion swarm with a high velocity and new electrons are continuously injected into the plasma region. Further, since this discharge takes place even at high vacua, the ratio of ions to neutral molecules is usually much higher than in gas-discharge plasmas. If it is assumed that no secondary electrons are released from electrodes bounding the plasma, it takes on a potential equal to or a few tenths of a volt less than the boundary electrode with lowest potential. This small potential difference between the electrode and the plasma serves as a shield for the ions in order to maintain the equilibrium[2] between ion production and leakage from the plasma. It is a function of the gas pressure.

In the practical case when the electrodes bounding the plasma region give off secondary electrons when struck by primary or by stray electrons the situation becomes somewhat different. Now a low-velocity electron stream having a high charge density is injected into the plasma region. Since most of these low-velocity secondaries do not contribute to ion production, they will tend to depress the plasma potential. This potential depression will hinder the loss of ions. As a consequence, since the tendencies are always in the direction of charge neutrality, additional secondaries can now enter the plasma. In actuality, of course, these various

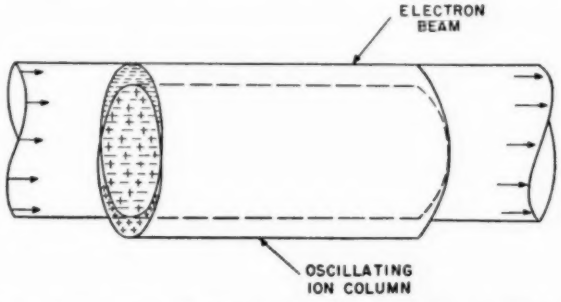


FIG. 2. Dipole mode of oscillation of cylindrical plasma ($g=1/\sqrt{2}$).

[9] K. G. Hernqvist, Proc. Inst. Radio Engrs. **39**, 1541 (1951).

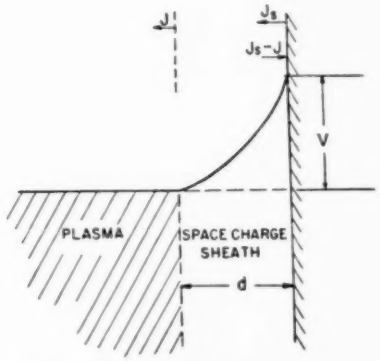


FIG. 3. Potential distribution at electron emitting plasma boundary electrode.

effects are simultaneous, the net result being an increase in plasma density. The secondary electron-current density J allowed to enter into the plasma in equilibrium will, for a certain structure and primary-electron flow, depend only on the ion production rate, that is, on the gas pressure. If it is assumed that the electrode bounding the plasma emits electrons with a Maxwellian velocity distribution (which may be an approximation to the secondary-electron distribution) and if the influence of the space charge of the primary-electron flow is neglected, then the conditions in the space-charge sheath are analogous to those between the cathode-surface and the virtual cathode in a diode. Thus the potential difference V between the boundary electrode and the plasma (see Fig. 3) and the sheath thickness d are determined by [10,11]

$$J=\eta\frac{V_w^{\frac{3}{2}}}{d^2} \tag{3}$$

$$J=J_s\exp\left[\frac{V}{V_0}\right] \tag{4}$$

where J_s is the saturation emission current density, V_w is the voltage corresponding to the maximum in the velocity distribution curve, $V_0=2V_w$ is the voltage corresponding to the average velocity of emitted electrons, and η is a function of J/J_s. For $J/J_s<1/30$ the value of η is approximately equal to 2.5×10^{-5}.

IV. MECHANISM FOR MAINTENANCE OF ION OSCILLATIONS IN A PLASMA BOUND BY AN ELECTRON EMITTING ELECTRODE

In Sec. I the natural frequencies of oscillation of an ion plasma have been discussed. If such oscillations are to be maintained, a mechanism is necessary by which energy is continuously transferred to the electric ac field associated with the ion vibrations. It has been shown[12] that such a mechanism exists for plasma

[10] I. Langmuir, Phys. Rev. **21**, 419 (1923).

[11] H. Rothe and W. Kleen, *Grundlagen und Kennlinien der Elektronenröhren* (Akademische Verlagsgesellschaft, Book Review Department, Leipzig, Germany, 1943), p. 33.

[12] G. Wehner, J. Appl. Phys. **21**, 62 (1950).

electron oscillations based on electron transit-time effects. Since the low frequency of plasma ion oscillations would require impractically low electron velocities for transit-time interaction, some other mechanism of energy transfer must be sought.

By reference to Fig. 3 and Eq. (3) it is seen that if the ion plasma oscillates in such a way as to vary the space-charge sheath thickness d, a modulation of the amount of secondary electron current J passing into the plasma field can occur. This modulated current will then interact with the ac electric plasma field associated with the ion oscillations. An accurate treatment of this interaction would require knowledge of damping as well as other amplitude limiting factors. However, using the simplified picture leading to Eqs. (3) and (4), it is seen that an increase in J would be expected when the ion plasma moves toward the boundary electrode. Simultaneously this electron stream will meet a retarding field within the plasma. On the contrary, as the ion plasma moves away from the boundary electrode, J would decrease and the plasma field be accelerating. Thus a net energy transfer may be expected from the modulated secondary emission current to the plasma field.

V. EXPERIMENTAL INVESTIGATION OF THE FREQUENCY DEPENDENCE OF PLASMA ION OSCILLATIONS IN ELECTRON BEAMS

Figure 4 shows schematically the experimental tube and measurement setup. The tube, which is an ion-trapping electron gun described elsewhere,[2] has spherically shaped cathode C and accelerating grid G. These project a conically formed electron beam through the aperture A to the collector F. If the aperture is made small enough so that some of the beam current is collected on A, ion-plasma oscillations of frequency f_R will be excited in the region between G and A, and the ac electric field setup will deflect the primary electron beam leading to an ac current partition between A and the collector F. Thus if a receiver is connected across a small resistor in series with F, a signal of frequency f_R depending on the density of the electron beam can be observed. In order to investigate the geometrical factors associated with these self-sustained oscillations a ring-formed electrode S was placed around the electron beam close to the aperture. The distance between S and A was about the same as the aperture diameter (1 mm), and no beam current was intercepted by S. An external ac voltage could be applied between S and A from a signal generator of variable frequency f_S.

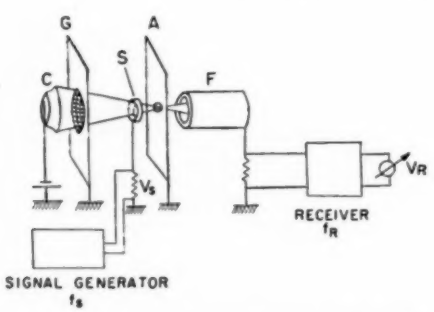


FIG. 4. Schematic drawing of ion-trapping gun and circuit for studying ion plasma oscillations.

An experiment was performed in which the frequency of the self-sustained oscillations was kept at a fixed value f_R and in which the receiver was constantly tuned to this frequency. The amplitude of the output voltage from the receiver V_R was then observed as the frequency f_S of the external voltage applied between S and A was varied. Figure 5 shows schematically the relation between V_R and f_S. No change in the amplitude of the self-sustained oscillations V_{R0} was observed until f_S was equal to f_R, at which point an increase in V_R was observed. If f_S was increased further to a value of f_x a decrease in V_R could be observed. Figure 6 shows f_x measured for different values of the frequency of the self-sustained oscillations equal to f_R. It is seen that for the region over which the self-sustained oscillations could be detected the relation $f_x=\sqrt{2}f_R$ is followed quite accurately. An explanation for this effect can be had if it is assumed that the self-sustained oscillations are of the dipole mode type (Fig. 2) with a geometry factor $g=1/\sqrt{2}$, being excited due to asymmetry between the electron beam and the aperture hole. Oscillations of the plane parallel type with a frequency $f_x=\sqrt{2}f_R$ are then excited by the external signal applied to S and tend to suppress the self-sustained oscillations. For certain well-aligned guns self-sustained oscillations of these two modes could be observed simultaneously, the dipole mode being, however, always greater in amplitude. This dipole mode of oscillation has been observed experimentally for plasma electron oscillations.[13]

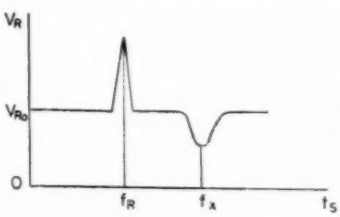


FIG. 5. Output amplitude of receiver (tuned to a fixed frequency f_R) *versus* frequency of signal applied to ring-formed electrode.

In order to determine the nature of the oscillating ions, small amounts of helium gas were introduced into the continuously pumped tube, the pressure being held below 10^{-5} mm Hg. The self-sustained oscillations were found to change to a frequency approximately 2.5 times higher when helium was inserted. Since the molecular weight of He is 4, this would, according to Eq. (2), correspond to a molecular weight of about $4\times2.5^2=25$ for the ions oscillating in the high-vacuum tube. Considering the low accuracy of this experiment the oscillating ions can be identified as nitrogen ions with $M=28$.

[13] D. Romell, Nature **167**, 243 (1951).

Measurement of the variation of frequency with ion density has been carried out for the ion-trapping gun.[2] These data are repeated here in Fig. 7, where relation (2) is also shown for $g=1/\sqrt{2}$ and $M=28$. It is seen that the measured densities are too high. This can be accounted for by the fact that the current density was measured at the aperture and the plasma density is very inhomogeneous. Therefore, since probably a considerable part of this inhomogeneous plasma is excited, the frequency determining density can be expected to be less than that measured as the electron density at the aperture. In fact, calculations of the plasma geometry show that the density $\frac{1}{2}$ millimeter away from the aperture would fit the theoretical curve very well.

VI. INVESTIGATIONS OF THE ENERGY EXCHANGE MECHANISM

Since a complete theory for the energy transfer to the plasma field can hardly be worked out for the complicated structures involved, only qualitative observations were made. Two ways are possible, in principle, to verify the mechanism proposed in Sec. IV. These are: (1) to influence the secondary electron current responsible for the oscillation excitation or (2) to inject a low-velocity electron flow into the plasma region from a separate electron source. Both of these methods were tried. The results fit the mechanism proposed, as far as the source of energy is concerned.

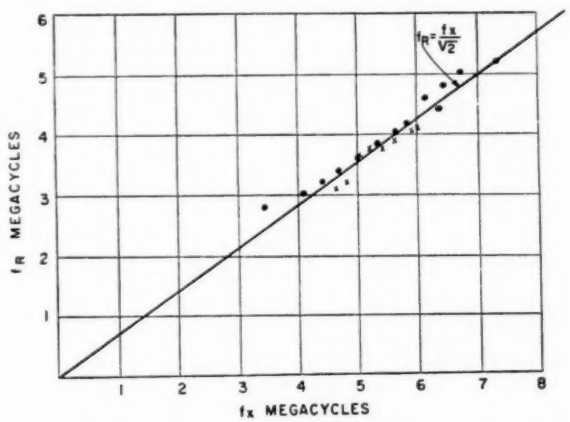


FIG. 6. Frequency f_R of self-sustained oscillations *versus* frequency f_x of excited oscillations.

The first method was used in connection with the ion-trapping gun (Fig. 4). The oscillations are here excited by the secondary electron current from the region around the aperture in a similar but more complicated way than described in Sec. IV. The secondary electron current allowed to enter the plasma region (corresponding to J) can be increased by increasing the gas pressure and would be expected to lead to an increased oscillation amplitude. An increase in the oscillation amplitude by a factor of 8.4 was

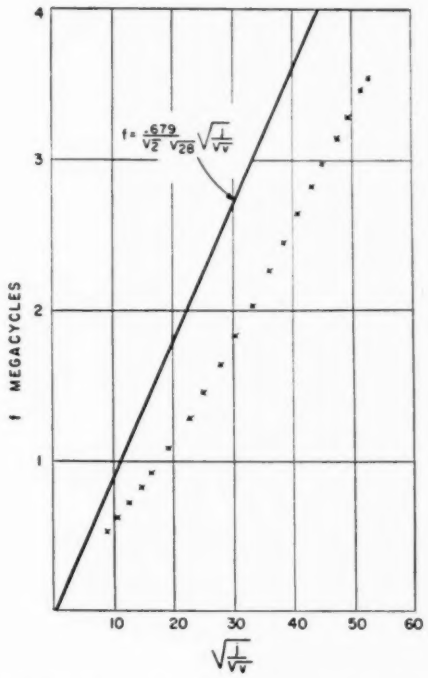


FIG. 7. Ion plasma frequency *versus* electron density.

observed while increasing the gas pressure from 4.5×10^{-6} to 6.3×10^{-5} mm Hg. J can also be influenced by increasing the potential of A with respect to G. Since the plasma always remains at the lowest boundary electrode potential,[14] the secondary electrons from A can be completely prevented from entering the plasma region. Experiments show that the oscillations are stopped if A is made 5 to 10 volts higher in potential

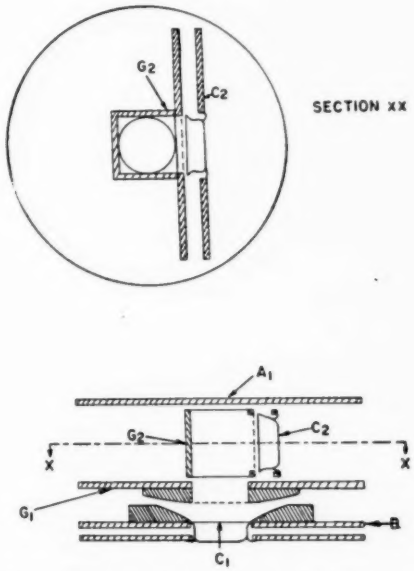


FIG. 8. Scale drawing of tube for excitation of plasma ion oscillations by an injected electron stream.

[14] E. G. Linder, Phys. Rev. **80**, 100 (1950).

than G. However, making A 50 or 100 volts lower than G has no appreciable effect on the amplitude of oscillation. The oscillations observed in tetrodes[9] behave similarly for changes in the anode potential with respect to the screen grid potential.

Instead of using the secondary electrons for oscillation excitation, a low-velocity electron stream injected at the plasma boundary could serve the same purpose. Figure 8 shows a scale drawing of a tube built for studying this effect. A Pierce gun structure having a flat oxide-coated cathode disk C_1, beam forming electrode B, and accelerating grid G_1 injects a parallel electron beam through the square tube G_2 to the anode A_1. The tubing G_2 has one side in the form of a mesh grid through which an electron flow can be injected from the cathode C_2.

In an experiment G_1 and G_2 were kept at a potential of 150 volts with respect to C_1. Keeping the potential of A_1 slightly higher completely prevented any self-sustained plasma oscillations to occur in the plasma formed with G_2 for reasons discussed above. Into this plasma a low-velocity electron stream was injected by applying a potential to C_2 of 10 volts or less with respect to G_2. This was found to excite oscillations in the plasma within G_2 in a way analogous to secondary electrons from a boundary electrode.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Anisotropic Diffusion Lengths in Diffusion Theory

BERNARD I. SPINRAD
Argonne National Laboratory, Lemont, Illinois
(Received September 25, 1954)

A set of formulas defining diffusion lengths in heterogeneous assemblies is presented. Even in the homogeneous limit, these formulas lead to anisotropic diffusion lengths for systems such as rod or slab assemblies which are microscopically anisotropic.

WITHIN the framework of diffusion theory, it is possible, with systems which are macroscopically anisotropic (such as rod or plate lattices), to observe anisotropies in the neutron equations of state. One such anisotropy is that neutron currents can decay in space at different rates depending on their orientations relative to preferred axes of the system.

A different approach to the same problem has been formulated by Dr. F. B. Estabrook (manuscript in preparation); in his approach an anisotropic migration length was deduced for multiplying systems consisting of finite parallel slabs by considering the critical equation of the finite system. As in this article, the assumption was made that the diffusion process takes place independently in the two media.

Formulas estimating the decay constant $1/L$, for neutron currents in a lattice are herewith presented for:

1. Currents parallel to a plate.
2. Currents perpendicular to a plate. A discussion of the problems involved in studying rod lattices is also included.

These studies all rest on the definition of L by an asymptotic experiment. In some cases, other idealized experiments may be used to define L for specific reactor calculations; the results given here are nevertheless significant for computing anisotropies.

(1) DIRECTIONS PARALLEL TO A PLATE

Let us consider a system of plates consisting of two materials (denoted by I and II), each with known thickness t, diffusion coefficient D, and inverse diffusion length κ. The materials are stacked alternately in an infinite array of slabs of infinite cross-sectional area. Diffusion theory and diffusion boundary conditions are assumed to hold at interfaces, and a valid solution of the homogeneous diffusion equation

$$D(\nabla^2\phi-\kappa^2\phi)=0 \tag{1}$$

is desired such that:

(a) $\partial\phi/\partial z=0$ at the center of the plates (z being the direction normal to the plates).

(b) $\lim_{\rho\to\infty}\phi$ is finite (ρ being the radius in the plane parallel to the plates relative to some fixed origin).

To determine the parallel diffusion length, $L_{\|}$, we place a source at $\rho=0$, and examine the asymptotic decay of the flux.

With an allowed source at the origin and condition (b), we may specify a solution as:

$$\phi_\kappa=f_\kappa(z)K_0(\kappa\rho), \tag{2}$$

where κ is some eigenvalue of the problem determined as will be described. An arbitrary source at the origin ($\rho=0$) may be obtained by summing ϕ over all ϕ_κ with assignable coefficients. The asymptotic decay is determined by the lowest value of κ, and this may be interpreted as defining

$$L_{\|}\equiv 1/\kappa. \tag{3}$$

To determine κ, let subscripts 1 and 2 be used to denote regions I and II. In region I, a solution of the

diffusion Eq. (1) in the form of Eq. (2) subject to boundary condition (*a*) is

$$f_{1\kappa}(z)=A\cos(\kappa^2-\kappa_1^2)^{\frac{1}{2}}z \tag{4a}$$

where A is an arbitrary constant and z is reckoned from the center of slab I. In region II,

$$f_{2\kappa}(z')=B\cos(\kappa^2-\kappa_2^2)^{\frac{1}{2}}z' \tag{4b}$$

B being an arbitrary constant. Here, z' has its origin at the center of slab II. If in (4a) or (4b), $\kappa^2<\kappa_i^2$, these are still solutions if the imaginary roots are substituted formally.

Then, under equality of flux and normal current at the interface ($z=t_1/2$, $z'=-t_2/2$),

$$A\cos(\kappa^2-\kappa_1^2)^{\frac{1}{2}}\frac{t_1}{2}=B\cos(\kappa^2-\kappa_2^2)^{\frac{1}{2}}\frac{t_2}{2} \tag{5a}$$

$$D_1(\kappa^2-\kappa_1^2)^{\frac{1}{2}}A\sin(\kappa^2-\kappa_1^2)^{\frac{1}{2}}\frac{t_1}{2}$$
$$=-D_2(\kappa^2-\kappa_2^2)^{\frac{1}{2}}B\sin(\kappa^2-\kappa_2^2)^{\frac{1}{2}}\frac{t_2}{2}. \tag{5b}$$

Hence,

$$D_1(\kappa^2-\kappa_1^2)^{\frac{1}{2}}\tan(\kappa^2-\kappa_1^2)^{\frac{1}{2}}\frac{t_1}{2}$$
$$=-D_2(\kappa^2-\kappa_2^2)^{\frac{1}{2}}\tan(\kappa^2-\kappa_2^2)^{\frac{1}{2}}\frac{t_2}{2}. \tag{6}$$

Equation (6) is the critical equation for κ. We expect at least one κ to lie between κ_1 and κ_2. A solution with $\kappa>\kappa_1$, κ_2 must have at least one of the tangents negative and, hence, the flux must have a root within the appropriate region; and if $\kappa<\kappa_1$, κ_2, both tangents convert to tanh, which is never negative for positive arguments. Thus, the lowest κ lies between κ_1 and κ_2 where κ_1 is now assumed to be the smaller. Then, finally,

$$D_1(\kappa^2-\kappa_1^2)^{\frac{1}{2}}\tan(\kappa^2-\kappa_1^2)^{\frac{1}{2}}\frac{t_1}{2}$$
$$=D_2(\kappa_2^2-\kappa^2)^{\frac{1}{2}}\tanh(\kappa_2^2-\kappa^2)^{\frac{1}{2}}\frac{t_2}{2}. \tag{7}$$

If the smallest solution of (7) is substituted into (3), a numerical value can be obtained for L_{II}.

For t_1, t_2 small or $\kappa_2^2-\kappa_1^2\ll\kappa_1^2$, we have almost homogeneous systems. Under these conditions, (7) may be reduced to

$$\kappa^2\doteq\frac{\kappa_1^2D_1t_1+\kappa_2^2D_2t_2}{D_1t_1+D_2t_2}. \tag{8a}$$

Using

$$\textstyle\sum_{ai}\equiv\kappa_i^2D_i \tag{9}$$

(8a) becomes

$$\kappa^2\doteq(\textstyle\sum_{a1}t_1+\sum_{a2}t_2)/(D_1t_1+D_2t_2) \tag{8b}$$

which shows that, in the limit, κ^2 approaches the volume weighted mean absorption cross section divided by the volume weighted mean diffusion coefficient. If the D's are markedly different, this is *not* the same limit as would be obtained for a uniformly homogeneous (as opposed to this stratified-homogeneous) system.

(2) DIRECTION PERPENDICULAR TO A PLATE

In this case, we take the same system as before, but consider fluxes constant in ρ, arising from plane sources at some z. Condition (a) must then be replaced by

$$\lim_{z\to\infty}\phi \text{ is finite.} \tag{8c}$$

Since, in diffusing from a source plane, neutrons must cross alternately regions I and II, we cannot now expect a solution of form $e^{-\kappa z}$ as might be assumed from analogy with the previous problem. Rather, we expect to obtain a solution such that equivalent points in adjacent cells will have equal flux/current ratios. Thus, if we specify as a cell a slab of region I to the left, followed by a slab of region II, then we expect the flux-to-current ratios at the two ends of the cell to be identical, and the flux at the right-hand edge to be less than that at the left (assuming a source somewhere to the left of our cell); for if the flux at the right were equal to or greater than the flux at the left, the source would be multiplying as the neutrons travel out from it.

Now let $\begin{bmatrix}\phi\\ J\end{bmatrix}$ be a vector describing neutron flux ϕ and negative current $J\equiv D\phi'$ everywhere in the system. The solution of the diffusion equation in region I may now be expressed as:

$$\begin{bmatrix}\phi_1\\ J_1\end{bmatrix}_z=\begin{bmatrix}\cosh\kappa_1 z & \dfrac{\sinh\kappa_1 z}{\kappa_1 D_1}\\ \kappa_1 D_1\sinh\kappa_1 z & \cosh\kappa_1 z\end{bmatrix}\begin{bmatrix}\phi_1\\ J_1\end{bmatrix}_{z=0}, \tag{10a}$$

and in region II as

$$\begin{bmatrix}\phi_2\\ J_2\end{bmatrix}_z=\begin{bmatrix}\cosh\kappa_2(z-t_1) & \dfrac{\sinh\kappa_1(z-t_1)}{\kappa_2 D_2}\\ \kappa_2 D_2\sinh\kappa_2(z-t_1) & \cosh\kappa_2(z-t_1)\end{bmatrix}\begin{bmatrix}\phi_2\\ J_2\end{bmatrix}_{z=t_1}. \tag{10b}$$

Since ϕ, J are continuous across the interface at t_1,

$$\begin{bmatrix}\phi\\ J\end{bmatrix}_{z=t_1+t_2}=\begin{bmatrix}\cosh\kappa_2 t_2 & \dfrac{\sinh\kappa_2 t_2}{\kappa_2 D_2}\\ \kappa_2 D_2\sinh\kappa_2 t_2 & \cosh\kappa_2 t_2\end{bmatrix}\times$$
$$\times\begin{bmatrix}\cosh\kappa_1 t_1 & \dfrac{\sinh\kappa_1 t_1}{\kappa_2 D_2}\\ \kappa_1 D_1\sinh\kappa_1 t_1 & \cosh\kappa_1 t_1\end{bmatrix}\begin{bmatrix}\phi\\ J\end{bmatrix}_{z=0}. \tag{11}$$

Our condition may be expressed, however, as:

$$\begin{bmatrix}\phi\\ J\end{bmatrix}_{z=t_1+t_2}=\lambda\begin{bmatrix}\phi\\ J\end{bmatrix}_{t=0} \tag{12}$$

so that the problem is reduced to finding an eigenvalue $\lambda<1$. We then identify λ with $\exp[-\kappa_\perp(t_1+t_2)]$ and obtain

$$L_\perp=1/\kappa_\perp. \tag{13}$$

To derive λ, we multiply the two matrices on the right of (11) and determine the eigenvalues of the combined matrix by subtracting λ from the diagonal terms and setting the determinant to zero. Thus:

$$\begin{vmatrix} \cosh\kappa_1 t_1\cosh\kappa_2 t_2+\dfrac{\kappa_1 D_1}{\kappa_2 D_2}\sinh\kappa_1 t_1\sinh\kappa_2 t_2-\lambda, & \dfrac{\sinh\kappa_1 t_1\cosh\kappa_2 t_2}{\kappa_1 D_1}+\dfrac{\sinh\kappa_2 t_2\cosh\kappa_1 t_1}{\kappa_2 D_2} \\ \kappa_1 D_1\sinh\kappa_1 t_1\cosh\kappa_2 t_2+\kappa_2 D_2\sinh\kappa_2 t_2\cosh\kappa_1 t_1, & \cosh\kappa_1 t_1\cosh\kappa_2 t_2+\dfrac{\kappa_2 D_2}{\kappa_1 D_1}\sinh\kappa_1 t_1\sinh\kappa_2 t_2-\lambda \end{vmatrix}=0 \tag{14}$$

or, reducing,

$$\lambda^2-2\lambda\left[\cosh\kappa_1 t_1\cosh\kappa_2 t_2+\left[\frac{\kappa_1 D_1}{\kappa_2 D_2}+\frac{\kappa_2 D_2}{\kappa_1 D_1}\right]\sinh\kappa_1 t_1\sinh\kappa_2 t_2\right]+1=0. \tag{15}$$

We note that the product of the two roots of λ is unity. This is what we expect for eigenfunctions which decay either toward the left or the right: that is, we may write

$$\lambda=\exp[\kappa_\perp(t_1+t_2)] \tag{16}$$

and obtain $\pm\kappa_\perp$ as solutions.

With (16) as a substitute, we may divide (15) by λ and obtain:

$$\cosh\kappa_\perp(t_1+t_2)=\cosh\kappa_1 t_1\cosh\kappa_2 t_2+\left[\frac{\kappa_1 D_1}{\kappa_2 D_2}+\frac{\kappa_2 D_2}{\kappa_1 D_1}\right]\sinh\kappa_1 t_1\sinh\kappa_2 t_2. \tag{17}$$

Equation (17) may be simplified in a number of ways, of which only one is presented here. If t_1, t_2 are small (homogeneous limit), we may expand both sides of (17) to terms in t^2. After some simplification, we get

$$\kappa_\perp^2=[(\textstyle\sum_{a_1}t_1+\sum_{a2}t_2)/(t_1+t_2)]\,[(t_1/D_1+t_2/D_2)/(t_1+t_2)]. \tag{18}$$

Thus, in the homogeneous limit, $\kappa_\perp^2$ approaches the true homogeneous κ^2, in which the $\sum_a$ and $1/D(\equiv 3\sum_{tr})$ are separately averaged.

(3) ROD LATTICES

The two pertinent directions of neutron current to consider for rod lattices are along the rod axis and across the rod array. The formulation of the relaxation length problem is sensitive to the rod lattice pattern and will not be reproduced here.

In the parallel direction, the problem of the circular rod in a circular cell may be explicitly set up by the method indicated for the slab. If, then, the homogeneous limit is taken, it is found that κ_{II}^2 is given by the volume-averaged macroscopic absorption cross section divided by the volume-averaged diffusion coefficient, as in the parallel slab problem. Further examination will verify that this homogeneous limit is not sensitive to rod or cell shape.

Across the rod array, the solution of a diffusion length problem is sufficiently complicated so that computing devices probably ought to be used for any specific system. A sample problem of square rods in square cells yielded, as the homogenous limit for $\kappa_\perp^2$, the volume-averaged macroscopic absorption cross section divided by an effective diffusion coefficient which is obtained by (1) finding $\bar{D}$ as for a parallel plate problem in that part of the cell where the current traverses both regions in parallel; and (2) averaging this $\bar{D}$ with the D of the outside region as in a perpendicular plate problem. No proof that this technique is extensible to other rod and cell shapes has been found, however, although the method proposed would seem physically reasonable.

(4) CONCLUSIONS

From the foregoing, the following qualitative conclusions can be stated:

(a) The effective diffusion length in a heterogeneous assembly is not a single number, but must rather be defined as a quantity which depends on the direction of the appropriate neutron current.

(b) When a homogeneous system is approached by scaling the heterogeneous lattice structure to arbitrarily small size, the anisotropy in diffusion length persists; that is, the diffusion length is sensitive to microscopic grain of a system.

Asymptotic Spherical Shock Decay

L. M. TANNENWALD*
U. S. Naval Ordnance Test Station, Inyokern, China Lake, California
(Received September 3, 1954)

It is shown that the description of viscous and heat conductive effects in asymptotic spherical shock decay can be reduced to the solution of an ordinary first-order differential equation relating shock strength and radius. The derivation of this equation is based, to a considerable extent, on a well-justified quasi-stationary approximation. Computational work required for handling specific situations is outlined.

I. INTRODUCTION

ALL parts of a spherical (gas) disturbance, propagating at sufficiently large distances r from the center, were considered until rather recently[1–6] to travel with ambient sound velocity and decay according to an r^{-1} law. The authors referred to,[1–6] however, noted that important cumulative effects arise from small, non-acoustic contributions to the local velocity of signal propagation. In particular, several of them[2–5] found that the shock strength (defined as the ratio of shock crest overpressure to ambient pressure) in a gas at a distance R falls faster than R^{-1} by a factor

$$(\ln R)^{-\frac{1}{2}}. \qquad (1.0)$$

The equations of motion of a perfect fluid and the Hugoniot shock conditions were used in deriving Eq. (1.0). These equations, unlike the Hugoniot conditions,[7] do not allow for the effects of viscosity or heat conduction. They are, however, still useful everywhere except in the immediate neighborhood of the shock (or shock-forming waves which are not considered in this paper). The magntiude of the accumulated error arising from this neglect of viscosity and heat conduction during long decay times is unknown. Therefore, it seems desirable to investigate the problem of weak spherical shock decay in a viscous and heat conducting medium (air) over large distances far from the origin of the disturbance.

In Sec. II of this paper a quasi-stationary treatment of the shock front phenomenon is justified for conditions that are usually satisfied in practice. Using this conclusion and various results gleaned from already existing works, an ordinary first-order differential equation relating shock strength and shock radius is derived next. The solution of this equation used in conjunction with the results of isentropic flow theory, as developed by Bethe and Whitham, results in a complete temporal description of the flow field. The previously mentioned technique appears much simpler than a treatment based on the direct integration of the parabolic system of three nonlinear partial differential equations of the problem. Section III contains an outline of numerical work that is required for use of the method developed in Sec. II; IBM work done in this connection is discussed. In conclusion, Sec. IV contains remarks on the scope of the method and its ease of application.

II. THEORY

The method of describing the flow field, given in this section, essentially owes its simplicity to three conditions.

(a) The shock front phenomenon is of a quasi-stationary nature. In other words, decaying and steady-state shock fronts of equal strength are very similar within a certain neighborhood of the shock front.

(b) The region where dissipative effects are important is extremely narrow in comparison to the extension of the whole disturbance.

(c) The extension of the whole disturbance, in turn, is very small in comparison to its distance from the origin.

Condition (a) holds because a molecule traversing a shock can easily adjust itself to the new surroundings in a time which is extremely short in comparison to the time it takes for a marked change to take place in the physical characteristics of a shock. Condition (b) is generally satisfied until the disturbance takes on a dominantly diffusive character. At that stage of decay, the problem ceases to be of practical importance. The problem is stipulated to allow condition (c) to be easily satisfied.

From (b) and (c) we conclude that the shock front may be considered to be plane to a high approximation. Thus, the results of Becker's[7] theory of the plane, steady-state shock in a viscous and a heat-conducting medium are usable there. The Becker theory, however, gives an over-all step-function behavior to the flow field variables; actually, they have roughly N wave profiles. Isentropic theory predicts N wave profiles, but it is less adequate than Becker's theory near the front of the wave. Thus, the best aspects of the two theories ought to be combined by means of an interpolation (see Fig. 1).

* Present address: Lockheed Aircraft Corporation, Missile Systems Division, Van Nuys, California.

[1] J. G. Kirkwood and H. A. Bethe, Office of Scientific Research and Development, OSRD No. 588 (1942).
[2] L. D. Landau, J. Phys. (U.S.S.R.) **6**, 229 (1942).
[3] DuMond, Cohen, Panofsky, and Deeds, J. Acoust. Soc. Am. **18**, 97 (1946).
[4] Bethe, *et al.*, U. S. Atomic Energy Commission, 2860 (declassified 1949).
[5] G. B. Whitham, Proc. Roy. Soc. (London) **A203**, 571 (1950).
[6] G. B. Whitham, Comm. Pure and Appl. Math. **5**, 301 (1952).
[7] R. Becker, Z. Physik **8**, 321 (1922).

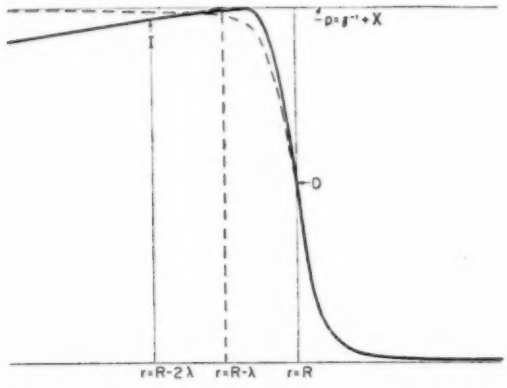


FIG. 1. The pressure distribution in the neighborhood of the shock front. The solid line exhibits the pressure variation in the isentropic, interpolation and dissipative regions which are characterized by $r \leqslant R-2\lambda$, $R-2\lambda \leqslant r \leqslant R$, and $r \geqslant R$, respectively. The broken lines show those portions of pressure profile shapes predicted by isentropic and dissipative theory that were not used in the theory presented.

It is shown in the latter part of this section that, unless unexpected cancellations arise, it is sufficient to consider the pressure-profile interpolation problem.

Denote the ambient density, temperature, and sound velocity by ρ_0, T_0, and a_0, respectively. The pressure, p, measured in units of $\rho_0 a_0^2$ is represented by the interpolation formula

$$f(p, r-R; \alpha^{(i)})=0 \tag{2.0}$$

between the dissipative and isentropic regimes. f is taken to be a linear function of certain adjustable constants $\alpha^{(i)}$; $i=0, 1, 2, \cdots$. All lengths are measured in units of the length, L_0, of the complete disturbance at time $t=0$.

The interpolation is stipulated by the following construction (Fig. 1). The dissipative solution is considered to describe the pressure drop of the shock front except within a distance λ of the pressure discontinuity given by isentropic theory. Further on in this section, λ is determined quantitatively. The pressure and its gradient are taken to be[7,8a]

$$p(D) \simeq \gamma^{-1}+X/2, \tag{2.1}$$

$$p_x(D) \simeq -(8\gamma)^{-1}(\gamma+1)\mathcal{R}X^2 \tag{2.2}$$

at the "dissipative" (D) junction point. γ, $X=X(R)$, μ, are respectively the ratio of specific heats, shock overpressure, viscosity, and $\mathcal{R}=\rho_0 a_0 L_0/4/3\mu$.

Throughout this article it will be taken for granted that $X^2 \ll 1$, and that reference to the use of this inequality is shown only by $\simeq$ signs in the appropriate relations. Equations (2.1) and (2.2) give the values of p and p_x at $x=0$[8b] for the steady-state (Prandtl number $=0.75$) pressure distribution, which has the following form: $p=\gamma^{-1}v^{-1}T$ with $v \simeq 1-X(1+\exp\Lambda^{-1}x)^{-1}$ and

$$T=1+2^{-1}(\gamma-1)[1+2^{-1}(\gamma+1)X][1-v^2],$$
$$\Lambda \simeq 2\gamma[(\gamma+1)\mathcal{R}X]^{-1}.$$

T and v are the temperature and specific volume in units of T_0 and ρ_0^{-1}, respectively.

The isentropic solution is considered adequate to describe the relatively gradual pressure drop behind the shock except within a distance λ of the pressure discontinuity. A simplified derivation of relations that are used below in connection with this theory is given in the Appendix. There it is shown that a disturbance which at time $t=0$ had the pressure profile $p=\gamma^{-1}+F(r)r^{-1}$ has at time (measured in units of $a_0^{-1}L_0$)t

$$p(I) \simeq \gamma^{-1}+X-\lambda p_r(I) \tag{2.3}$$

and

$$p_r(I) \simeq F_Y[R(1+2^{-1}(\gamma+1)F_Y \ln Y^{-1}R)]^{-1} \tag{2.4}$$

at the "isentropic" (I) junction. $r=Y$ is the location, at $t=0$, of the signal which reaches the I-junction at time t. Y is to be considered a function of R or (and) X. For example, if at $t=0$ an N wave of strength $\gamma\bar{X}$ has the form $p=\gamma^{-1}+2\bar{X}(r-\bar{R}+\frac{1}{2})$ for $\bar{R}-1 \leqslant r \leqslant \bar{R}$ and elsewhere $p=\gamma^{-1}$, then $Y=\bar{R}-2^{-1}[1-(\bar{R}/\bar{R})(X/\bar{X})]$ and $F_Y=+2\bar{R}\bar{X}$.

The maximum pressure lies somewhere between the I- and D-junction points. This requires, at least for some f, that a fifth condition should be satisfied at some x for which $-\lambda \leqslant x \leqslant -2\Lambda$, namely,

$$\text{and} \qquad \begin{matrix} p_x=0 \\ p=\gamma^{-1}+X. \end{matrix} \tag{2.5}$$

Relations (2.1)–(2.5) probably represent the minimum number of quantitative geometrical conditions necessary for a decent interpolation.

Somewhat better shock profiles can, in principle, be obtained by requiring Eq. (2.0) to satisfy continuity conditions for higher than first-order derivatives at the two junction points. In fact, the method for fitting together the pressure profile has been stipulated so as to obviate, to a considerable extent, the need for handling these conditions. This has been done by placing the junction points away from the well-localized region where $|p_{rr}|$ is expected to be large (see Fig. 1). At the D-junction, the value of $|p_{rr}|$ given by dissipative theory is smaller than the largest $|p_{rr}|$ expected in the interpolation interval, $|p''|$, by roughly a factor X. The ratio of $|p_{rr}|$ at the I-junction, derived from isentropic theory, to $|p''|$ is much smaller than X. Since we seek an approximate solution of a system of equations containing no derivatives of higher than second order, the use of conditions Eqs. (2.1)–(2.5) should be quite adequate in first approximation.

In addition to the geometrical conditions, it is necessary to fix λ, the thickness of the shock front, by

[8a] A subscript on any quantity signifies partial differentation of the quantity with respect to its subscript. The ambient state's $_0$ subscript is the only exception to this convention.

[8b] The origin of the x coordinate system is the point $r=R$. R is chosen to lie at the D junction. Therefore, the point $x=r-R=0$ moves outward with the instantaneous velocity of the shock.

physical reasoning. It is easy to see that λ is a monotonically decreasing function of $|p''|$. The latter, therefore, serves equally well as a measure of the shock thickness. For a given shock strength a very accurate estimate of p'', which is justified by the quasi-stationary nature of phenomenon at the shock front, is obtained from its (near) equality to the minimum p_{rr} of dissipative theory, $p''(\mathfrak{d})$. No restriction is otherwise made on p'' except the requirement that p'' lie well within the interpolation interval, just as $p''(\mathfrak{d})$ does.

In summary, it should be noted that the complete flow picture, for a given shock strength (or fixed time), can be found by fitting together known dissipative and isentropic solutions by means of a relation (2.0) having five $\alpha^{(i)}$ constants.

The method for determining the flow picture for all times is derived next. Consider four functions $\eta^{(j)}(j=0, 1, 2, 3)$ which satisfy an equation[9]

$$\eta_t{}^{(0)}+\eta^{(i)}{}_{x(i)}=0 \tag{2.6}$$

and are only functions of $x=r-R$ (with $r^2=x^{(i)2}$) and $\alpha^{(i)}=\alpha^{(i)}(R,X)$[10]—apart from a factor $(x^{(i)}/r)$ in each $\eta^{(i\neq 0)}$—within a spherical shell Ω whose inner and outer radii are, respectively, $R-2\lambda(R,X)$ and R. Then

$$\int_\Omega r^2\eta_t{}^{(0)}dr+[\![r^2[n^{(j)}\eta^{(j)}]]\!]=0. \tag{2.7}$$

In Eq. (2.7), $n^{(j)}$ denotes the $x^{(j)}$ component of the outward normal of Ω; for any function $\varphi(r)$, $[\![\varphi]\!]\equiv\varphi(R)-\varphi(R-2\lambda)$. If X is considered as $X=\Xi(t)$, then the integral in Eq. (2.7) may be transformed into

$$\int_\Omega r^2dr\{\eta^{(0)}{}_{\alpha(i)}[\alpha_R{}^{(i)}R_t+\alpha_X{}^{(i)}\Xi_t]+\eta_x{}^{(0)}(-R_t)\}$$
$$=R_t\int_\Omega r^2dr\{\eta^{(0)}{}_{\alpha(i)}[\alpha_R{}^{(i)}+\alpha_X{}^{(i)}X_R]-\eta_x{}^{(0)}\},$$

since $R_t=\Xi_tX_R^{-1}$. $R_t\equiv U(X)=[1+2^{-1}(\gamma+1)X]^{\frac{1}{2}}$ is the shock velocity.

The last integral above can be simplified by using $R-r\ll R$. Then,

$$\int_\Omega r^2dr\{\eta^{(0)}{}_{\alpha(i)}[\alpha_X{}^{(i)}X_R+\alpha_R{}^{(i)}]-\eta_x{}^{(0)}\}$$
$$\simeq R^2\int dr\{\eta^{(0)}{}_{\alpha(i)}[\alpha_X{}^{(i)}X_R+\alpha_R{}^{(i)}]-\eta_x{}^{(0)}\}$$
$$=R^2\left\{(\alpha_X{}^{(i)}X_R+\alpha_R{}^{(i)})\int_{-2\lambda}^{0}dx\eta^{(0)}{}_{\alpha(i)}-[\![\eta^{(0)}]\!]\right\}.$$

Similarly, $[\![r^2n^{(j)}\eta^{(j)}]\!]\equiv[\![r^2\eta]\!]\simeq R^2[\![\eta]\!]$. Thus Eq. (2.7) finally becomes

$$U[\alpha_X{}^{(i)}X_R+\alpha_R{}^{(i)}]\int_{-2\lambda}^{0}dx\eta^{(0)}{}_{\alpha(i)}+[\![\eta-U\eta^{(0)}]\!]=0 \tag{2.8}$$

which is an ordinary, first-order differential equation relating X and R. Equation (2.8) is put into more usable form by taking[11]

$$\eta^{(0)}=\frac{p}{(\gamma-1)}+\frac{\rho u^2}{2}$$
$$\eta^{(k\neq 0)}=pu^{(k)}-3/4\Re^{-1}\{u^{(i)}[u^{(k)}{}_{x(i)}+u^{(i)}{}_{x(k)}]$$
$$-2/3u^{(k)}u^{(i)}{}_{x(i)}\}-3/4(\gamma-1)^{-1}\Re^{-1}T_{x(k)}$$
$$+u^{(k)}\rho\left[\frac{T}{\gamma(\gamma-1)}+\frac{u^2}{2}\right]. \tag{2.9}$$

$u^{(k)}$ and ρ, denote the gas velocity in units of a_0 and density in units of ρ_0, respectively. The expressions (2.9) for $\eta^{(j)}$ can be simplified using $X\ll 1$. In particular, since $p_t[(\rho u^2)_t]^{-1}\sim X^{-1}$,

$$\eta^{(0)}\simeq(\gamma-1)^{-1}p. \tag{2.9'}$$

Equation (2.9′) now proves the assertion made earlier in this section that only the pressure interpolation $\alpha^{(i)}$'s must be considered in this problem; in general, the number of sets of $\alpha^{(i)}$ that must be dealt with is equal to the number of flow variables in $\eta^{(0)}$. Thus, although there are other independent flux equations having simpler $\eta^{(k\neq 0)}$ than Eq. (2.9) such as

$$\eta^{(0)}=\rho,$$
$$\eta^{(k\neq 0)}=\rho u^{(k)}, \tag{2.10}$$

little, if any, benefit is derived from their usage, because usually by far the largest practical difficulty lies in finding the $\alpha^{(i)}$ and λ. Choice (2.9′) for $\eta^{(0)}$ has the advantage of referring to the most easily measured flow variable.

The dependence of $\alpha^{(i)}$ and λ on R and X and, in turn, the solution $X(R)$ of Eq. (2.8) must be found by numerical means.

III. NUMERICAL WORK

In the practical application of Eq. (2.8), one first meets the problem of choosing f in Eq. (2.0). Many types of functions for f were investigated with respect to their ease of handling by straightforward algebra and IBM 701.

The quadratic forms

$$\alpha^{(-1)}p^2+\alpha^{(0)}px+\alpha^{(1)}p+\alpha^{(2)}x^2+\alpha^{(3)}x+\alpha^{(4)}=0, \tag{3.0}$$

where $\alpha^{(-1)}=1$ or 0, seemed by far the most tractable functions for use with conditions (2.1)–(2.5) and the one on p''. Note that the case of $\alpha^{(-1)}=0$ can be disregarded

[9] The double index summation convention (over index values $=1, 2, 3$) is used hereafter unless indications to the contrary are given.

[10] The $\alpha^{(i)}$ can here refer to interpolation constants of more than one flow variable.

[11] Van Deemter, Phys. Rev. **85**, 1049 (1952).

because the p'' condition ($p_{rrr}=0$) requires $\alpha^{(0)}=0$; this, in turn, has $\alpha^{(1)}=0$ as a consequence, and then the remaining $\alpha^{(i)}$'s are insufficient to handle all the interpolation conditions. Extensive calculations with the IBM 701 indicate that, even for $\alpha^{(-1)}=1$, there is no quadratic form which can satisfy the interpolation conditions for the numerical values of interest.[12]

The results of the computation imply, therefore, that even if the correct number of adjustable constants is available, relation (3.0) in particular and in general every f, may not always be able to satisfy the type of interpolation conditions (2.1)–(2.5).[13] The interpolation problem is in this respect distinctly more difficult than the often-met problem of finding a polynomial function which takes on given values at specified points. It also should be mentioned that the computational work is lengthier in the former problem because, in contrast to the latter, the adjustable constants must be found from a set of nonlinear equations. The solution of Eq. (2.8), which governs the whole shock decay, is a relatively minor task in comparison to the effort required to find a suitable f and its α's and λ. Straightforward numerical techniques can be used to solve Eq. (2.8) because it is an ordinary first-order differential equation.[14]

The writer hopes that it will be possible for him to return to this problem and give numerical details at a later time.

IV. CONCLUSIONS

By exploiting physical assumptions that are well justified, it has been possible to circumvent the intractability of the parabolic-type system of nonlinear differential equations that describe our shock decay problems. The method also can handle other time-dependent problems whose isentropic solutions are available. Some examples of these are the decay of a plane or cylindrical shock, and the back shock(s) of a spherical wave. Although it is usually only of academic interest, a still more accurate description of the shock front phenomena can be treated with our method by introducing better expressions for the steady-state dissipative shock.

After taking advantage of the quasi-stationary nature of the shock front phenomenon, the previously developed mathematical technique appears in retrospect to be similar in concept to the methods introduced by v. Kármán and Pohlhausen[15] for investigating boundary layer problems. In spite of the greater mathematical difficulty of the shock wave problem, the usefulness of a similar technique in the boundary layer problem is an indication of the power to be expected from the approach described in this paper.

ACKNOWLEDGMENT

This problem was suggested to the writer by H. A. Bethe. The benefit derived by the writer from clarifying discussions with H. A. Bethe and G. B. Whitham is gratefully acknowledged. The writer is indebted to R. P. Stein for having shared some of the computing labors.

APPENDIX

Some relations of isentropic theory which were stated in Sec. II are briefly derived here for the sake of completeness. The physical ideas and some of the results are essentially the same as those of Bethe.[4]

Use will be made of the adiabatic relation $a\simeq 1+(\gamma-1)/2\delta p$ between the local sound velocity, a (in units of a_0), and the sufficiently small overpressure, $\delta p(\equiv p-\gamma^{-1})$. At least locally, δp can be expected to satisfy the acoustic wave equation

$$p_{rr}+2p_r r^{-1}=p_{tt} \tag{5.0}$$

to a good approximation. Equation (5.0) has the well-known outgoing wave solution

$$\delta p=F(y)r^{-1} \tag{5.1}$$

where $y=r-t$. One can also obtain

$$u\simeq\delta p=F(y)r^{-1} \tag{5.2}$$

if δp from Eq. (5.1) is substituted into $u_t+p_r\simeq 0$.

It is to be noted that the aforementioned y is the location, at $t=0$, of a signal which propagates with velocity 1 and passes through r at time t. Thus, only for the very few wavelets of the disturbance for which $\delta p=0$ can Eqs. (5.1) and (5.2), with $y=r-t$, be strictly valid over arbitrarily large distances because the velocity of signal propagation is actually $a+u$. Nevertheless, one can still expect the geometrical attenuation to be given by the form of Eqs. (5.1) and (5.2) for all wavelets with $(\gamma^{-1}\gg)\delta p\neq 0$. In order to be in harmony with $a+u-1\neq 0$, y should be considered generally as the location, at $t=0$, of the signal which passed through r at time t. With this modified meaning for y, one can expect Eqs. (5.1) and (5.2) to be valid over very large distance ($r\gg 1$). Calculations of a more formal nature[5] confirm this conjecture.

[12] Probably, an investigation should be made next to see whether certain types of f, particularly polynomials, can satisfy condition (2.5) to an accuracy of order X by only requiring f to obey the following conditions: (a) (2.1)–(2.4), (b) $p''=p''(b)$, and (c) p_{rr} has only one relatively large minimum between $r=R$ and $r=R-2\lambda$. This approximation is suggested by $\lambda\ll 1$ and $|p_r(I)|\ll|p_x(D)|$.

[13] The correctness of this conclusion can be seen from the following example. Consider a situation where $p(I)$, $p(D)$, $p_r(I)p_x(D)$, and λ are fixed, and for which $p(I)>p(D)$, $|p_r(I)|<|p_x(D)|$, and $p_r(I)(p_x(D))^{-1}<0$; the maximum p, however, is allowed to vary. If the maximum p then becomes greater than a certain value, it will be necessary for the curve which can fit the interpolation conditions to have more than one extremum for p_{rr}. Thus Eq. (3.0), for example, cannot successfully interpolate in the situation formulated here (which is slightly different from the one occurring in the problem) because

$$p_{rr}=-\tfrac{1}{2}\frac{[\alpha^{(0)2}-4\alpha^{(2)}][\alpha^{(1)2}-4\alpha^{(4)}]-[\alpha^{(0)}\alpha^{(1)}-2\alpha^{(3)}]^2}{\{[\alpha^{(0)2}-4\alpha^{(2)}](r-R)^2+2[\alpha^{(0)}\alpha^{(1)}-2\alpha^{(3)}](r-R)+[\alpha^{(1)2}-4\alpha^{(2)}]\}^{\frac{1}{2}}}$$

[14] W. E. Milne, *Numerical Calculus* (Princeton University Press, Princeton, New Jersey, 1949).

[15] Schlichting, *Grenzschicht-Theorie* (G. Braun, Karlsruhe, 1951).

The arrival time of a signal at r is, therefore,

$$t=\int_y^r \frac{dr}{a+u}$$

$$=\int_y^r dr\left[1+\frac{(\gamma+1)}{2}F(y)r^{-1}\right]^{-1}$$

$$=r-y-\frac{(\gamma+1)}{2}F(y)\ln\left[y^{-1}r+\frac{(\gamma+1)}{2}F(y)y^{-1}\right]\times\left[1+\frac{(\gamma+1)}{2}F(y)y^{-1}\right]^{-1}$$

$$\simeq r-y-2^{-1}(\gamma+1)F(y)\ln y^{-1}r \qquad (5.3)$$

because $\delta p\ll\gamma^{-1}$.

Substitution of

$$y_r=-t_r t_y^{-1}\simeq[1+2^{-1}(\gamma+1)F_y\ln y^{-1}r]^{-1}$$

into $p_r=[F(y)r^{-1}]_r\simeq F_y y_r r^{-1}$ yields

$$p_r\simeq F_y(r[1+2^{-1}(\gamma+1)F_y\ln y^{-1}r])^{-1}. \qquad (5.4)$$

Relation (2.4) is obtained by putting $r\simeq R$ and $y=Y$ in Eq. (5.4). Relation (2.3) is derived from the first two terms of the Taylor series expansion for $p(R-\lambda,t)\simeq\gamma^{-1}+X$ about the I-junction point.

If, as in Sec. II, $p=\gamma^{-1}+2\bar{X}(y-\bar{R}+\frac{1}{2})$, etc., at $t=0$, then $F(y)\simeq 2\bar{X}\bar{R}(y-\bar{R}+\frac{1}{2})$, $F_Y\simeq 2\bar{X}\bar{R}$, $F(Y)R^{-1}=2\bar{X}(\bar{R}/R)(Y-\bar{R}+\frac{1}{2})=X$ and $Y=R-2^{-1}[1-(R/\bar{R})\times(X/\bar{X})]$.

The reader is referred to the cited work of Bethe[4] and Whitham[5,6] for additional details.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Propagation of Shock Waves in Aluminum

H. DEAN MALLORY
U. S. Naval Ordnance Laboratory, White Oak, Silver Spring, Maryland

(Received July 31, 1954)

The velocity of shock waves in aluminum and the associated translational motions, produced by metal-metal impact, have been determined by an electrical contact technique. The results obtained have been used to evaluate an equation of state for the metal.

INTRODUCTION

THE rate of propagation of shock waves in, and the free boundary velocities of metals are experimentally measurable quantities which can be used in the calculation of pressure behind the shock front. A study of the propagation of shocks induced by contact with a detonating explosive can thus in principle yield information relative to detonation pressures. The desirability of such a measure of explosive performance is obvious. Calculations relative to the initial fragment velocity of shells further requires consideration of the velocity of sound in compressed metals; a suitable equation of state is therefore essential.

An equation of state which requires the evaluation of two constants has been given by Pack, Evans, and James.[1] They have evaluated the constants by using Bridgman's compression data and applied it to lead and steel with some success. In the present paper the equation is applied to aluminum and some experimental verification is given.

Our initial experiments for measuring explosion pressures were similar to those now reported except that a bare explosive charge was in contact with the aluminum target plate. Although free surface velocities could be measured with some precision, shock wave velocity measurements were erratic. This was believed due to violent pressure changes in the reaction zone of the detonation wave, and therefore a method for producing shock waves in the aluminum targets was needed which was not subject to reaction zone effects. The impact of aluminum on aluminum was successfully used as shown by the data to be presented.

The impact velocities used were sufficiently great to insure that the shock waves produced were all in the range of stable shocks[2] where the shock wave velocity is at least as great as the velocity of the plane elastic wave in aluminum. Some as yet unpublished data show the velocity of this elastic wave in aluminum to be just under 6100 m/sec, whereas the lowest computed shock wave velocity associated with the present experiments is 6140 m/sec at the free surface of 1-inch thick targets. Computations have been carried out over the compressed aluminum density range 2.7 through 3.7 g/cc although in the low density region, up to about 3.05 g/cc, the corresponding velocities are subject to error due to the plane elastic wave which precedes the shock wave.

[1] Pack, Evans, and James, Proc. Phys. Soc. (London) **60**, 1–8 (1948).

[2] Walsh, Shreffler, and Willig, J. Appl. Phys. **24**, 349 (1953).

WAVE VELOCITIES

The equation for P given by Pack, Evans, and James[1] is

$$P=\alpha\left\{\exp\left(\beta\frac{\nu_0^{\frac{1}{3}}-\nu^{\frac{1}{3}}}{\nu_0^{\frac{1}{3}}}\right)-1\right\}\Big/\left(\frac{\nu}{\nu_0}\right)^{\frac{1}{3}}, \qquad (1)$$

which can be expanded by substituting a power series in P for $-\Delta\nu/\nu_0$. To the second degree in P one obtains:

$$-\frac{\Delta\nu}{\nu_0}=\frac{3}{\alpha\beta}P-\frac{3}{2\alpha^2\beta^2}(\beta+6)P^2.$$

Comparison with Bridgman's formula

$$-\frac{\Delta\nu}{\nu_0}=AP-BP^2,$$

where for aluminum[3] $A=13.696\times10^{-13}$ and $B=5.20\times10^{-24}$, permits evaluation of the constants α and β which are 20.603×10^{10} and 10.632, respectively, where all are in cgs units.

An expression for P in series form is

$$P=\alpha\left(\frac{\rho}{\rho_0}\right)^{\frac{1}{3}}\sum_{n=1}^{\infty}\frac{x^n}{\underline{n}} \quad \text{where} \quad x=\beta\left[1-\left(\frac{\rho_0}{\rho}\right)^{\frac{1}{3}}\right].$$

This can be differentiated with respect to ρ to yield the square of the sound speed in the absence of rigidity:

$$c^2=\frac{M}{3}\sum_{n=0}^{\infty}\frac{\beta^n}{\underline{n+1}}\frac{(\rho^{\frac{1}{3}}-\rho_0^{\frac{1}{3}})^n}{\rho^{n+2/3}}[2\rho^{\frac{1}{3}}+(n-1)\rho_0^{\frac{1}{3}}] \qquad (2)$$

where $M=\alpha\beta\rho_0^{-\frac{1}{3}}$, c is the sound speed, ρ_0 and ρ are initial and final densities, respectively, and α and β are the constants of Eq. (1).

Shock wave velocities may be computed by considering the hydrodynamic equations of continuity

$$\rho(D-u)=\rho_0 D$$

and momentum

$$P=\rho_0 Du, \qquad (3)$$

where u is the material velocity behind the shock front. From these two equations is obtained

$$D^2=\frac{\rho}{\rho_0}\left(\frac{P}{\rho-\rho_0}\right).$$

With the help of Eqs. (1), (2), and (3) shock velocities (D) in aluminum have been computed and compared with sound velocities (c) in Fig. 1.

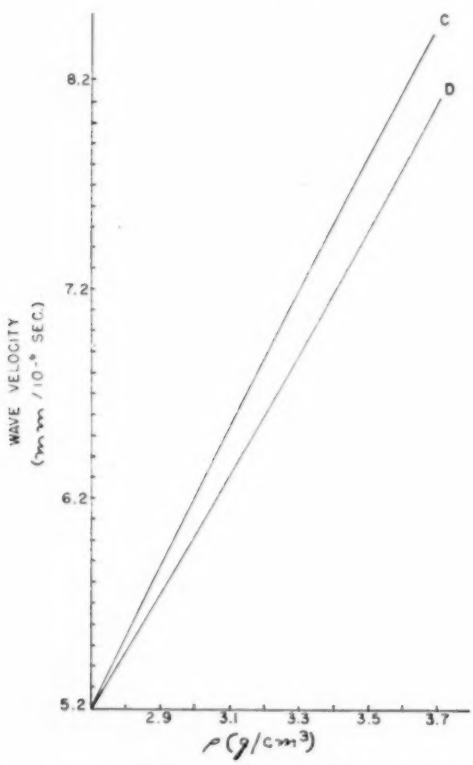


FIG. 1. Wave velocities in compressed aluminum as a function of aluminum density.

[3] P. W. Bridgman, *The Physics of High Pressure* (G. Bell and Sons, Ltd., London, England, 1949), p. 160.

THE VELOCITY OF SHOCKED METAL AT A FREE BOUNDARY[4]

We now suppose that a plane shock wave with velocity D moves into a metal at rest in a manner such that its front is parallel to a plane free boundary and inquire into the free boundary velocity on arrival of the shock. Since material is following the shock at a velocity u, and metal does not move off in a permanently compressed state, it is evident that unloading must occur at the free boundary and cause a wave of rarefaction to be sent into the compressed metal. It is necessary then that the free boundary has a velocity in excess of u. This excess velocity is the Riemann function σ where

$$\sigma=\int_{\rho}^{\rho_0}\frac{c\,d\rho}{\rho} \quad \text{and} \quad c=\left(\frac{dP}{d\rho}\right)_s^{\frac{1}{2}}.$$

Using Eq. (2), σ can easily be integrated numerically or in series form with somewhat more difficulty. A series solution for σ is given in Eq. (4)

$$\sigma=\left(\frac{3\alpha}{\beta\rho_0}\right)^{\frac{1}{3}}\int_0^x e^{x/2}\left[1+\sum 2^n\prod_{K=1}^{n}\times\left(\frac{2K-3}{2K}\right)\left(\frac{e^{-x}-1}{\beta-x}\right)^n\right]dx. \qquad (4)$$

This equation can be approximated by

$$\sigma=\left(\frac{3\alpha}{\beta\rho_0}\right)^{\frac{1}{3}}\int_0^x e^{x/2}\left(1+\frac{x}{7}\right)^{\frac{1}{3}}dx,$$

where the range of interest is $0<x<1$.

[4] Some discussion of this, as applied to waves in water, is given by Robert H. Cole, *Underwater Explosions* (Princeton University Press, Princeton, New Jersey 1948).

Values of σ and u have been calculated at corresponding aluminum densities and tabulated in Table I. According to Table I, values for u and σ are nearly equal over the density range covered by the experimental data, and the approximation

$$u_s = u + \sigma \doteq 2u$$

is correct to within 1 percent. In order to obtain these values of u and σ, we have assumed that differences between the Rankins-Hugoniot $P-V$ curve and the adiabatic curve for aluminum are small and that Eq. (1) is a close approximation to either.

EXPERIMENTAL DETAILS

In each test, a two pound explosive charge 3×3 inches at the face and 4 inches high was detonated in contact with an aluminum disk 6 inches in diameter and 1 inch thick which was separated from another similar disk, having one of several thicknesses, by a $\frac{3}{8}$-inch air gap. Measurements of shock and free surface velocity were made on the latter disk while the former served as the shock driver. The free surface velocity of the driver disk is a function of both the pressure developed by the explosive used and the disk thickness. Variation of these two parameters permits a wide selection of pressures incident on the target disk at impact. A change in driver disk thickness calls for a compensating change in the air gap, the purpose of which is to allow an unloading wave to proceed into the driver a distance sufficient to insure no complicating interactions with the shock wave developed on impact with the target.

The scheme used for measuring the velocities was similar to that used by Pack, Evans, and James[1] although 16 electrical contact pins were used for each shot instead of a single one. Eight of these were insulated and set into holes drilled to various depths in the target to be shorted out on passage of the shock waves, while the remaining eight were spaced outside the disk in such a manner that the moving free surface would contact each in turn. The pin contactors were arranged in two concentric circles, one of $\frac{5}{8}$ inches and the other 1 inch in diameter at the target center. The impacting detonation wave was made plane by using a composite booster as described by Cook,[5] ahead of the main charge. A plane impact on the target disk was desirable because of the circular pin geometry. Electrical pulses made by contact of the grounded moving disk with the charged pins were led through 500 ft of coaxial cable to an oscilloscope where the scope beam deflections were photographed by a still camera. Synchronization with the oscilloscope was achieved by also sending the pulse from the first pin through a second shorter piece of cable in a manner such that it started the sweep of the scope beam in time to record itself and the others as they came through the long cable. The oscilloscope had a 5–10 mc frequency response and about a 2 microsecond time base, single sweep.

TABLE I. Properties of aluminum under shock.

ρ (g/cm³)	P (dynes/cm²) ×10⁻¹⁰	D (m/sec)	c (m/sec)	u (m/sec)	σ (m/sec)
2.7	0	(5200)	5200	0	0
2.8	2.886	5471	5540	195	195
2.9	6.146	5745	5877	396	396
3.0	9.793	6023	6209	602	600
3.1	13.85	6305	6540	814	809
3.2	18.36	6597	6869	1031	1022
3.3	23.31	6891	7197	1253	1238
3.4	28.72	7181	7520	1481	1458
3.5	34.62	7490	7842	1712	1680
3.6	41.03	7797	8162	1949	1906
3.7	47.97	8108	8481	2191	2134

TABLE II. Least square values of the shock velocity constants.

Constants	Least square values of the constants ± probable error ¼-in. targets	1-in. targets
a_0	-0.0017 ± 0.0054	-0.0115 ± 0.0056
a_1	0.1381 ± 0.0046	0.1425 ± 0.0012
a_2	$9.328\times10^{-4}\pm8.03\times10^{-4}$	$1.888\times10^{-4}\pm5.28\times10^{-5}$

Internal (shock) pins were set in holes drilled into a target disk with their tips insulated by mica 0.0007 inch thick. A time interval determined from an internal pin consists of the real transit time of the shock wave through the metal plus the delay required for the hole bottom to deflect and drive the pin through its insulating mica wafer. In the case of free boundaries, our data show the boundary velocity decreasing with increasing target thickness. Mica backed boundaries should behave similarly and therefore the delay time for hole bottom deflection should increase as the shock wave penetrates further into a target. No exact knowledge of this type of boundary motion is available; however, from the data in Table II one can say that if the insulation were air instead of mica, a maximum correction of only 10^{-8} second would be required. Due to the smallness of the expected correction it was ignored.

External pins were insulated at their tips by an oil film insignificantly thick while the shafts were coated with glyptal resin and spaghetti tubing. A schematic representation of the experiment is shown in Fig. 2.

TREATMENT OF THE DATA AND RESULTS

It is known from various other explosive loading experiments, in addition to those described here, that neither the free boundary velocity as a function of free flight distance nor the shock velocity as a function of target penetration is quite linear. We have therefore assumed a relation giving time as a quadratic function of distance for both types of data.

The circular pin contactor geometry requires that target motion be one-dimensional over the interval of measurement. For the thin targets this requirement was expected to be satisfied, but for the thick targets, peripheral rarefaction effects were expected to distort

[5] J. H. Cook, Research (London) 1, 474–477 (1948).

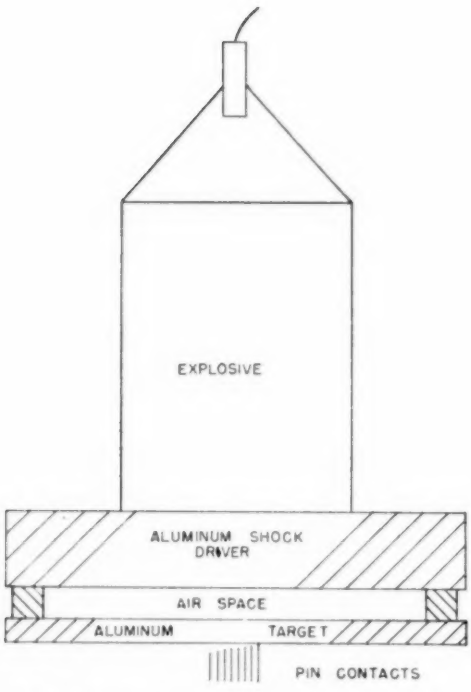


FIG. 2. Pin contactor and target arrangement.

the planeness of the wave. Subsequent wave shape experiments with a rotating-mirror streak camera have shown a free surface curvature of 0.1 microsecond, center to edge, over the $\frac{5}{8}$-inch center portion of the 1-inch thick targets. Since the plane wave boosters used in these experiments produce a detonation wave which is plane to within 0.05 microsecond, it was evident that some distortion was present. The wave arriving at the free surface of the $\frac{1}{2}$-inch thick targets was still plane to within 0.05 microsecond over the center $\frac{5}{8}$-inch diameter and therefore free surface motion was one-dimensional to within the limits of the experiment.

Edge rarefaction effects, then, may conceivably have some influence on the 1-inch target data but the following considerations indicate that this is not serious: (1) the free surface velocity data for the $\frac{1}{2}$- and 1-inch targets, falls on the curve computed from the velocities observed for "thin" targets, (2) free surface velocity data for 1-inch targets, using a 1-inch diameter pin contactor circle, showed much more scattered than comparable data from a $\frac{5}{8}$-inch diameter circle. Data from the smaller pin circle, though somewhat less reliable than that with thin targets, was about normal for the experiment.

Measurements of shock wave transit times were made on $\frac{1}{4}$- and 1-inch thick targets and the data fitted by least squares to the equation

$$t=a_0+a_1l+a_2l^2, \tag{5}$$

where l (mm) is the depth of penetration of aluminum by the shock wave and t is the time in microseconds. The least square values of the constants along with their probable errors are listed in Table II. Values for the constants in Table II are not directly comparable for the following reason: A number 1 pin extended completely through a target so that its tip was flush with the impacted surface. It is quite probable that the air shock wave formed ahead of a driver plate was sufficiently ionized to cause electrical contact before metal-metal impact occurred. Distance measurements were therefore started from the number 2 pins. Initial shock wave velocities are calculated from Eq. (5) by obtaining dl/dt and using a negative distance for l (the position of pin 1 relative to the zero position of pin 2). The mean pin 1 position for eight 1″ targets was -3.604 mm while that for seven $\frac{1}{4}$″ targets was -0.813 mm.

Free boundary data were fitted by least squares to the equation

$$t=b_0+b_1x+b_2x^2, \tag{6}$$

where t(microseconds) is the transit time of the moving aluminum surface from pin to pin, and x is the free flight travel distance in millimeters. Values for the constants are listed in Table III.

If now the u *vs* D data in Table I is fitted to the equation

$$u=m_0+m_1D+m_2D^2 \tag{7}$$

over the expected range of aluminum compression $3\leqslant\rho\leqslant3.4$ g/cc, we obtain an equation which can be used to predict, say, the surface velocity at any thickness if it is known for two. Differentiation of Eq. (5) to obtain $D=dl/dt$ and substitution into Eq. (7) yields

$$u=-3.0762+\frac{0.48591}{a_1+2a_2l}+\frac{0.020720}{(a_1+2a_2l)^2}. \tag{8}$$

The initial free boundary velocities ($u_s\doteq2u$) of the driver plate alone (zero target thickness) and the $\frac{1}{4}$ inch targets were considered to be the most reliable and were used in Eq. (8) to obtain a_1 and a_2 independent of the shock velocity measurements. They were 0.13999 and 4.4360×10^{-4}, respectively. On this basis

TABLE III. Least square values of the free boundary velocity constants for various target thickness.

Constant	Least square value of the constant ± probable error				
	Driver plate alone 0 in. target	$\frac{1}{4}$ in.	$\frac{1}{2}$ in.	$\frac{3}{4}$ in.	1 in.
b_0	0.0024±0.0037	0.0132±0.0074	−0.0046±0.0012	−0.0044±0.0068	−0.0263±0.0114
b_1	0.3439±0.0069	0.3610±0.0139	0.4041±0.0013	0.4754±0.0127	0.7254±0.0210
b_2	$(5.424\pm2.682)\times10^{-3}$	$(12.48\pm5.37)\times10^{-3}$	$(2.207\pm0.600)\times10^{-3}$	$(9.447\pm4.869)\times10^{-3}$	$(-3.298\pm8.000)\times10^{-3}$

we predict $u_s = 1396$ m/sec at $l = 25.4$ mm. The experimental velocity was 1379 ± 40 m/sec. The agreement is somewhat better than could reasonably have been expected in view of the long interval of extrapolation (see Fig. 3).

The free boundary velocity is more sensitive to pressure changes than is the shock wave velocity and in addition is more convenient to measure with precision. We have therefore regarded values of u_s as the most important. However, reasonable agreement exists between Eq. (8) and the experimental evaluation of Eq. (5) especially for the initial shock wave velocity. Here, at zero penetration, the neglect of time corrections due to the free boundary velocity decrease with penetration distance is completely negligible and the shock velocity as calculated from both sets of data can be weighted according to probable errors and a significant value obtained. From the 1-in. data we find 7089 ± 63 m/sec and from the $\frac{1}{4}$-in. data 7322 ± 257. The weighted mean is then 7130 ± 63 which compares well with the computed value of 7143 m/sec.

At a distance of 1 in. from the impact boundary the weighted mean of the experimental shock velocity is nearly 10 percent too low. This is attributed to a slowing down of the edges of the plane shock wave by peripheral rarefaction waves which have significantly distorted the flow ahead of the shock velocity measuring pins which were set in the 1-inch diameter circle. As previously mentioned, flow ahead of the free surface velocity measuring pins grouped in a $\frac{5}{8}$-inch diameter circle, was still essentially plane. This indicates that data from targets of thickness greater than 1 inch cannot be used under the size limitations of the experiment. The fact that peripheral rarefactions have lowered the shock wave velocity at some distance from the impact boundary does not cause serious error in the value of this quantity computed for zero target thickness. This results from the use of a quadratic observation equation which follows curvature in the time-distance data without regard to the cause.

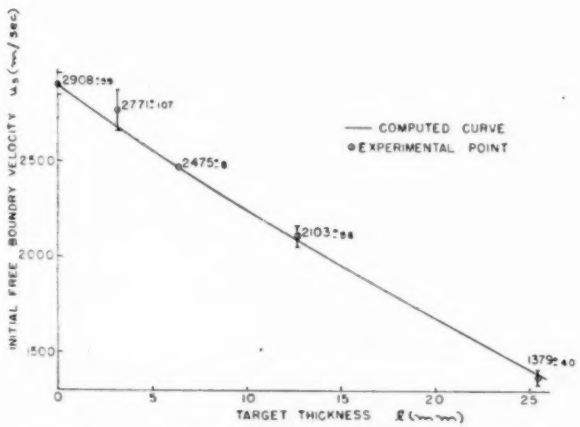


FIG. 3. Initial free boundary velocity of aluminum targets as a function of target thickness.

TABLE IV. Probable errors of the observation equations by external and internal consistency for all data.

Target type	Number of shots	Probable error (μsec) of the observation equations $R_{external}$	$R_{internal}$	R_{ext}/R_{int}
1-in. shock	8	0.0024	0.0065	2.7
$\frac{1}{4}$-in. shock	7	0.0024	0.0113	4.7
1-in. surface	6	0.0047	0.0230	4.9
$\frac{1}{2}$-in. surface	6	0.0029	0.0059	2.0
$\frac{1}{4}$-in. surface	8	0.0020	0.0042	2.1
$\frac{1}{8}$-in. surface	4	0.0030	0.0106	3.5
0-in. surface	6	0.0015	0.0030	2.0

All probable errors previously quoted are based on internal consistency since they are believed to be the more significant. However, following Birge,[6] we have also computed the probable errors (in time) of the observation equations by external consistency. These are given in Table IV. The ratio of the two probable errors $R_{external}/R_{internal}$ is a test for systematic errors in the data. Except for statistical fluctuations the ratio should be unity, although the data may be considered satisfactory if the ratio is less than 5.

WAVES IN STEEL

While doing some exploratory work involving direct detonation wave impact on steel targets it was noticed that if the internal pins were pressed in tightly, so their cellophane insulators were partially penetrated, the velocity obtained was about that reported by Pack, Evans, and James[1] but otherwise somewhat lower than their computed shock. Possibly their ball contact was sufficiently near the target surface to be shorted out by the small deflection of the metal surface on passage of the low amplitude elastic wave preceding the shock in steel.

CONCLUSIONS

The data presented indicate that Eq. (1) is applicable to aluminum in addition to steel and lead as reported by Pack, Evans, and James. By using the high velocity impact of metal on a metal target, a shock wave can be produced in the target. The shock wave so produced has a measurable velocity of propagation which is a smoothly decreasing function of target thickness.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to the following persons without whose help this work could not have been done. Dr. S. J. Jacobs for introducing the problem and solving many problems attending the electrical measurements, Mr. James Counihan, Mr. Joseph Schleicher, Jr., and Mr. Halcom Curtis for the experimental work, Miss Mary Fix for the numerical calculations, and Mr. Nathaniel Coleburn for the work relative to Eq. (4).

[6] Raymond T. Birge, Phys. Rev. **40**, 207 (1932).

Noise in One-Dimensional Electron Beams*

HERMANN A. HAUS
Research Laboratory of Electronics and Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received September 9, 1954)

A generalized theory of noise in one-dimensional electron beams is developed with the aid of the theory of four-terminal networks. No specific assumptions are made as to the input noise velocity and current modulations and their correlation at the potential minimum. The ensuing theory is simpler and more general than the corresponding theories used in the past. Transformations are discussed of the standing wave of the mean-square noise current in drift regions by means of "lossless beam transducers." It is shown that such transformations can be reduced to the formalism of conventional impedance transformations. Expressions are derived for the minimum obtainable noise figures of a velocity-jump amplifier, a klystron, and a traveling-wave tube. It is proved that these expressions present an ultimate limit for the noise figures of the respective tubes with regard to transformations by beam transducers with and without loss.

1. INTRODUCTION

THE design of low noise amplifiers of the beam type requires an understanding of the propagation of noise in electron beams. The theory of noise that has been used most widely for this purpose is based on Pierce's work.[1] Three assumptions are made in Pierce's analysis:

(a) The electron beam is treated as a one-dimensional medium.

(b) It is assumed that the random velocity fluctuations of the electrons can be represented by an equivalent model of a beam in which all electrons pass any given cross section of the beam with the same velocity fluctuation superimposed upon the time average velocity of the beam.

(c) The noise input conditions are assumed at the potential minimum in terms of Rack's equivalent noise velocity modulation.[2]

The first assumption seems necessary if simplicity of the mathematical formalism is desired. The implications of the second assumption, the single-velocity assumption, have been the subject of a paper by D. A. Watkins,[3] and unpublished work by A. M. Clogston, L. R. Walker, and the author. The results of this work indicated that the single-velocity approach to the problem of noise fluctuations in an electron beam is a good approximation to the actual multivelocity flow, subject to one condition: the spread in velocity of the electrons at any cross section in the beam must be small compared to their average velocity.

The most debatable assumption seems to be the third assumption concerning the noise input conditions at the potential minimum. Watkins[4] supplemented the assumed velocity modulation at the potential minimum with a current modulation corresponding to pure shot noise. He assumed that the velocity and current modulations at the potential minimum are uncorrelated. However, there now exists no rigorous high-frequency analysis of the noise process in the region of the potential minimum which would take into account both the space-charge forces and the velocity distribution of the electrons. In the absence of such an analysis, the magnitude of the current and velocity modulation at the potential minimum and the amount of correlation between them can be fixed only by hypothesis. Further, the average velocity of the electrons at the potential minimum is comparable to their velocity spread. The conditions are violated under which the validity of the single-velocity approximation has been proved. The noise excitation of the electron flow in this region cannot be characterized safely in terms of a velocity and current modulation.

The analysis presented in this paper retains the assumptions of a one-dimensional single-velocity flow. The input conditions are stated, however, at a cross section beyond the potential minimum at which the conditions are met for an approximate single-velocity analysis of the noise problem. Obviously, the input conditions at this reference plane are not known any better than for the conventional noise analysis discussed above. Therefore, the input conditions are formulated in the most general possible form. A formalism ensues which is self-consistent and of greater generality than the conventional noise analysis. The region of the beam in which the single-velocity approximation is invalid is excluded from the investigation. The greater generality of the approach leads to a simplicity not achieved by the conventional analysis. General problems that were not heretofore treated can be solved.

2. A SECTION OF A BEAM AS A PASSIVE FOUR-TERMINAL NETWORK

It is advantageous to introduce the terminology of four-terminal network theory into the theory of one-

* This paper is based on an excerpt from a doctoral thesis in the Department of Electrical Engineering, Massachusetts Institute of Technology. The work was supported in part by the U. S. Signal Corps; the Office of Scientific Research, Air Research and Development Command; and the Office of Naval Research.

[1] J. R. Pierce, *Traveling-Wave Tubes* (D. Van Nostrand Company, Inc., New York, 1950).
[2] A. J. Rack, Bell System Tech. J. **17**, 592 (1938).
[3] D. A. Watkins, J. Appl. Phys. **23**, 568 (1952).
[4] D. A. Watkins, thesis, Rept. 31, Electronics Research Laboratory, Stanford University, Stanford, California (1951).

dimensional electron beams. A section of a one-dimensional beam that passes through several time average accelerations and decelerations is shown in the schematic of Fig. 1. It is assumed that a single-velocity analysis of the beam beyond the cross section at z_0 is valid. By the assumptions of small signal theory, the velocity and current modulations at any reference cross section z_1 determine the velocity and current modulations at a later cross section z_2 in terms of linear relations. Instead of the complex velocity modulation at a cross section z, $v_1(z)$, we shall use the beam voltage modulation, $V_1(z)$, defined by L. J. Chu[5]:

$$V_1(z)=-\frac{m}{e}u(z)v_1(z), \tag{1}$$

where e/m is the charge-to-mass ratio defined positive and $u(z)$ is the average velocity of the beam at the cross section z. The relations between the complex amplitude of the convection current modulation, $I_1(z)$, and the beam voltage modulation $V_1(z)$ at two cross sections of the beam can be written in the form

$$\begin{aligned} V_1(z_2)&=A(\omega)V_1(z_1)+B(\omega)I_1(z_1)\\ I_1(z_2)&=C(\omega)V_1(z_1)+D(\omega)I_1(z_1). \end{aligned} \tag{2}$$

Equations (2) have the form of four-terminal network equations. The coefficients $A(\omega)$, $B(\omega)$, $C(\omega)$, and $D(\omega)$ are functions of the time average potential distribution along the beam between the cross sections z_1 and z_2. They do not satisfy, in general, the reciprocity relation $A(\omega)D(\omega)-B(\omega)C(\omega)=1$. We shall refer to a section of a beam between two reference cross sections as a "beam transducer."

In analogy to network theory, a power can be defined by[5]

$$\tfrac{1}{2}\,\mathrm{Re}[V_1(z)I_1^*(z)]. \tag{3}$$

The words "kinetic power" have been coined by L. J. Chu to denote expression (3). If electromagnetic power is extracted from the beam in the region between the cross sections z_1 and z_2, the kinetic power at z_2 is less than that at z_1. The corresponding beam transducer is "lossy." If no electromagnetic power is extracted in the region between z_1 and z_2, we have

$$\mathrm{Re}[V_1(z_1)I_1^*(z_1)]=\mathrm{Re}[V_1(z_2)I_1^*(z_2)].$$

The corresponding beam transducer is "lossless." In this particular case, Eqs. (2) have the form

$$\begin{aligned} V_1(z_2)&=[A_0(\omega)V_1(z_1)+jB_0(\omega)I_1(z_1)]e^{j\phi}\\ I_1(z_2)&=[-jC_0(\omega)V_1(z_1)+D_0(\omega)I_1(z_1)]e^{j\phi} \end{aligned} \tag{4}$$

where the coefficients A_0, B_0, C_0, and D_0 are real and satisfy the relation

$$A_0(\omega)D_0(\omega)-B_0(\omega)C_0(\omega)=1 \tag{4a}$$

and ϕ is some phase angle.[6] Velocity jumps, the accelerating region in an electron gun between a plane beyond the potential minimum and the anode, are lossless beam transducers.

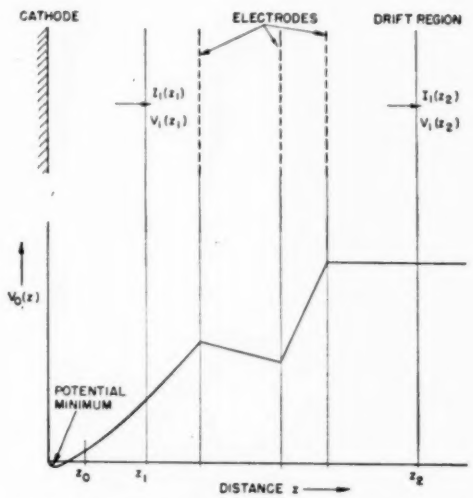


FIG. 1. Schematic of a one-dimensional beam and its time average potential distribution, $V_0(z)$.

3. FORMULATION OF THE NOISE PROBLEM

The noise in the electron beam is assumed to be a stationary statistical process. A statistical process can be defined only in terms of time or ensemble averages. The auto-correlation functions of the noise voltage and current modulations, and the cross correlation function between the noise voltage and current modulations, define the noise process sufficiently for our purpose.[7] We proceed to find the relation between the auto-correlation function of the noise voltage modulation at the output in terms of the auto- and cross-correlation functions of the voltage and current modulations at the input of a beam transducer with the matrix

$$\begin{bmatrix} A(\omega) & B(\omega) \\ C(\omega) & D(\omega) \end{bmatrix}.$$

It is convenient to transform Eqs. (2) for this purpose into the time domain. They assume the form

$$V_1(z_2,t)=\int_{-\infty}^{\infty}a(\nu)V_1(z_1,t-\nu)d\nu+\int_{-\infty}^{\infty}b(\nu)I_1(z_1,t-\nu)d\nu, \tag{5a}$$

$$I_1(z_2,t)=\int_{-\infty}^{\infty}c(\nu)V_1(z_1,t-\nu)d\nu+\int_{-\infty}^{\infty}d(\nu)I_1(z_1,t-\nu)d\nu. \tag{5b}$$

[5] L. J. Chu (unpublished notes).

[6] H. A. Haus, J. Appl. Phys. **25**, 1500 (1954).

[7] Y. W. Lee, "Application of statistical methods to communication problems," Tech. Rept. No. 181, Research Laboratory of Electronics, Massachusetts Institute of Technology (1950).

The kernels of the integrals are Fourier transforms of the complex coefficients $A(\omega)$, $B(\omega)$, $C(\omega)$, and $D(\omega)$; e.g.,

$$a(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}A(\omega)e^{j\omega t}d\omega$$

and

$$A(\omega)=\int_{-\infty}^{\infty}a(t)e^{-j\omega t}dt. \tag{6}$$

The unit impulse response functions of a physical system, that is, $a(t)$, $b(t)$, $c(t)$, and $d(t)$, satisfy the condition $a(t)=0$ for $t<0$, etc. The auto-correlation function of the noise voltage modulation at the cross section z_2, $\phi(z_2,\tau)$, is found from Eqs. (5) by a correlation operation (the details are given in Appendix I).

$$\begin{aligned}\phi(z_2,\tau)=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}a(\nu)a(\sigma)\phi(z_1,\tau+\nu-\sigma)d\nu d\sigma\\&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}b(\nu)b(\sigma)\psi(z_1,\tau+\nu-\sigma)d\nu d\sigma\\&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}a(\sigma)b(\nu)[\theta_{iv}(z_1,\tau+\nu-\sigma)\\&\qquad+\theta_{iv}(z_1,-\tau+\nu-\sigma)]d\nu d\sigma\end{aligned} \tag{7}$$

where

$$\psi(z_1,\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}I_1(z_1,t)I_1(z_1,t+\tau)dt$$

is the auto-correlation function of the noise current modulation, and

$$\theta_{iv}(z_1,\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}I_1(z_1,t)V_1(z_1,t+\tau)dt$$

is the cross-correlation function between the current and voltage modulations. The self-power density spectrum (SPDS) of the voltage modulation, $\Phi(z,\omega)$, is the Fourier transform of the auto-correlation function[7]

$$\Phi(z,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\phi(z,\tau)e^{-j\omega\tau}d\tau.$$

A Fourier transformation of Eq. (7), whose details are given in Appendix II, gives

$$\begin{aligned}\Phi(z_2,\omega)=&|A(\omega)|^2\Phi(z_1,\omega)+|B(\omega)|^2\Psi(z_1,\omega)\\&+A(\omega)B^*(\omega)\Theta_{iv}(z_1,\omega)+A^*(\omega)B(\omega)\Theta_{iv}{}^*(z_1,\omega)\end{aligned} \tag{8}$$

where $\Psi(z_1,\omega)$ is the SPDS of the noise current modulation, and $\Theta_{iv}(z,\omega)$ is the cross power density spectrum (CPDS) between the noise current and voltage.

$$\Theta_{iv}(z,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\theta_{iv}(z,\tau)e^{-j\omega\tau}d\tau.$$

Similarly, one can prove the following relations among the power density spectra:

$$\begin{aligned}\Psi(z_2,\omega)=&|C(\omega)|^2\Phi(z_1,\omega)+|D(\omega)|^2\Psi(z_1,\omega)\\&+C(\omega)D^*(\omega)\Theta_{iv}(z_1,\omega)+C^*(\omega)D(\omega)\Theta_{iv}{}^*(z_1,\omega)\end{aligned} \tag{9}$$

$$\begin{aligned}\Theta_{iv}(z_2,\omega)=&A(\omega)C^*(\omega)\Phi(z_1,\omega)+B(\omega)D^*(\omega)\Psi(z_1,\omega)\\&+A(\omega)D^*(\omega)\Theta_{iv}(z_1,\omega)+B(\omega)C^*(\omega)\Theta_{iv}{}^*(z_1,\omega).\end{aligned} \tag{10}$$

In the literature of noise in electron beams the magnitude of the noise current and velocity modulations are given in terms of the mean-square fluctuations per unit frequency. These quantities are related to the corresponding power density spectra by a factor of 4π. Thus, for instance, the SPDS of a pure shot noise current is $\Psi(\omega)=eI_0/2\pi$.

4. THE STANDING-WAVE PATTERN OF NOISE CURRENT IN A DRIFT REGION

The electromagnetic structure of a klystron or a traveling-wave tube in which the electron beam interacts with the electromagnetic fields is usually preceded by a drift region. The noise figure of these tubes depends upon the standing waves of the SPDS's of the noise voltage and current in the beam in the drift region. We find the expressions for the standing waves in terms of generalized input conditions.

Equations (8), (9), and (10) are valid for any beam transducer. A drift region between two cross sections z_1 and z_2 is a beam transducer with the matrix elements

$$\begin{aligned}A(\omega)&=\cos\theta e^{-j\omega T} & B(\omega)&=-jZ_0\sin\theta e^{-j\omega T}\\C(\omega)&=-jY_0\sin\theta e^{-j\omega T} & D(\omega)&=\cos\theta e^{-j\omega T},\end{aligned} \tag{11}$$

where ωT is the transit angle of the electrons between the two reference cross sections; θ is the transit angle measured in terms of the plasma frequency, $\omega_p=[(e/m)\times(|\rho_0|/\epsilon)]^{\frac{1}{2}}$ (ρ_0 is the time average space-charge density of the electrons); Z_0, a negative real quantity, is the characteristic impedance of the beam defined by

$$Z_0=\frac{1}{Y_0}=-2\frac{V_0}{I_0}\frac{\omega_p}{\omega},$$

where V_0 is the time average potential, and I_0 the time average current of the beam.

If a beam of finite cross section is considered, an analysis of the noise propagation in the lowest order space-charge mode is possible along similar lines. The plasma frequency ω_p has to be replaced in the expressions given above by the reduced plasma frequency of the mode.[4]

We find the power density spectra in terms of the input power density spectra at the cross section z_1 as a function of the length of the drift region θ by the use of Eqs. (8), (9), (10), and (11).

$$\begin{aligned}\Phi(z_2,\omega)=&\cos^2\theta\Phi(z_1,\omega)\\&+Z_0{}^2\sin^2\theta\Psi(z_1,\omega)-Z_0\sin 2\theta\Lambda(z_1,\omega),\end{aligned} \tag{12}$$

where $\Lambda(z,\omega)$ is the imaginary part of the CPDS $\Theta_{iv}(z,\omega)$

$$\Lambda(z,\omega)=\mathrm{Im}[\Theta_{iv}(z,\omega)] \tag{13}$$

$$\Psi(z_2,\omega)=Y_0^2\sin^2\theta\Phi(z_1,\omega)+\cos^2\theta\Psi(z_1,\omega)+Y_0\sin 2\theta\Lambda(z_1,\omega) \tag{14}$$

$$\Lambda(z_2,\omega)=\tfrac{1}{2}Z_0[Y_0^2\Phi(z_1,\omega)-\Psi(z_1,\omega)]\sin 2\theta+\cos 2\theta\Lambda(z_1,\omega). \tag{15}$$

Introducing a symbol for the real part of the CPDS, $\Theta_{iv}(z,\omega)$,

$$\mathrm{Re}[\Theta_{iv}(z,\omega)]=\Pi(z,\omega) \tag{16}$$

we find for this quantity the relation $\Pi(z_2,\omega)=\Pi(z_1,\omega)$. The conservation of the real part of the CPDS, $\Pi(z,\omega)$, is a consequence of the conservation of kinetic power in a drift region, which is a lossless beam transducer.

Equations (12) and (14) show that the SPDS's of the noise voltage and current as functions of distance along the drift region have the form of standing waves shown in Fig. 2. The maxima of $\Phi(z,\omega)$ and $\Psi(z,\omega)$ lie spatially a quarter of a plasma wavelength apart. The imaginary part of the CPDS, $\Lambda(z,\omega)$, is zero at the position of a maximum or minimum of $\Phi(z,\omega)$ and $\Psi(z,\omega)$. The shape of the standing-wave pattern is fixed entirely by the values of the SPDS's, $\Phi(z_1,\omega)$ and $\Psi(z_1,\omega)$, and the imaginary part of the CPDS, $\Lambda(z_1,\omega)$, at one cross section along the beam. Conversely, it is true that the knowledge of the standing-wave pattern of the SPDS's, $\Phi(z,\omega)$ and $\Psi(z,\omega)$, contains the information about the values of $\Phi(z_1,\omega)$, $\Psi(z_1,\omega)$, and $\Lambda(z_1,\omega)$ at a cross section z_1, but does not give an indication as to the value of the real part of the CPDS, $\Pi(z,\omega)$. It is shown in Appendix III that the product of the maximum of the SPDS of the current, Ψ_{max}, and its minimum, Ψ_{min}, is given by

$$Z_0^2\Psi_{max}\Psi_{min}=[\Phi(z_1,\omega)\Psi(z_1,\omega)-\Lambda^2(z_1,\omega)]. \tag{17}$$

The expression on the right-hand side of Eq. (17) is, in general, *not* equal to the square of the real part of the CPDS, II. It is known from correlation theory that the product of the SPDS's of two quantities is greater than, or at least equal to, the square of the absolute magnitude of their CPDS;

$$\Phi(z,\omega)\Psi(z,\omega)\geq|\Theta_{iv}(z,\omega)|^2=\Pi^2(z,\omega)+\Lambda^2(z,\omega)$$

or

$$\Phi(z,\omega)\Psi(z,\omega)-\Lambda^2(z,\omega)\geq\Pi^2(z,\omega). \tag{18}$$

The equality sign in relation (18) holds when a definite phase can be defined between the voltage and current in the frequency range, ω, $\omega+d\omega$. Let us call such a process "coherent process."

When a coherent process and a general noise process have the same values of $\Phi(z_1,\omega)$, $\Psi(z_1,\omega)$, and $\Lambda(z_1,\omega)$ at the cross section z_1, the standing-wave patterns of the SPDS's from either of the two processes are the same. We shall make use of this important fact later.

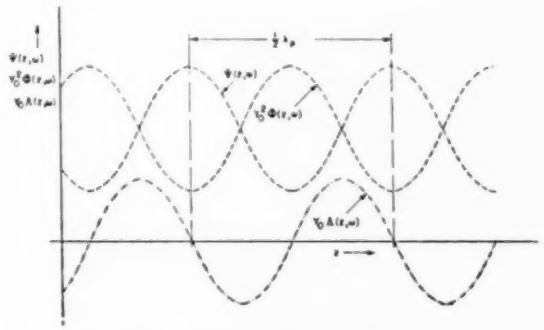


FIG. 2. The standing-wave patterns of the power density spectra in a drift region (λ_p is the plasma wavelength).

5. TRANSFORMATIONS OF THE NOISE BY MEANS OF LOSSLESS BEAM TRANSDUCERS

The standing-wave patterns of the SPDS's, $\Phi(z,\omega)$ and $\Psi(z,\omega)$, in a drift region are determined partly by the emission process from the cathode and by the interaction of the electrons in the region of the potential minimum up to the cross section z_0 (see Fig. 1); and partly by the interaction of the electrons in the region between the cross section z_0 and the beginning of the drift region. The cross section z_0 is so chosen that the single velocity approximation is valid at and beyond z_0. The region between the cross section z_0 and the drift region is a beam transducer. Most frequently lossless transducers are used.

In this section we shall determine the extent to which the noise in the drift region is determined by the input conditions at the cross section z_0 independent of the particular lossless beam transducer used. Further, the problem of the noise transformations by means of lossless transducers will be reduced to the simpler problem of conventional impedance transformations of lossless four-terminal networks.

Let the coordinate z_1 mark the beginning of the drift region. The transformations of the power density spectra at the cross sections z_0 and z_1 are found by the use of Eqs. (8), (9), and (10). The particular form of the elements of a matrix for a lossless beam transducer as shown in Eq. (4) is taken into account. One obtains

$$\Phi(z_1,\omega)=A_0^2\Phi(z_0,\omega)+B_0^2\Psi(z_0,\omega)+2A_0(\omega)B_0(\omega)\Lambda(z_0,\omega), \tag{19}$$

$$\Psi(z_1,\omega)=C_0^2(\omega)\Phi(z_0,\omega)+D_0^2(\omega)\Psi(z_0,\omega)+2C_0(\omega)D_0(\omega)\Lambda(z_0,\omega), \tag{20}$$

$$\Theta_{iv}(z_1,\omega)=jA_0(\omega)C_0(\omega)\Phi(z_0,\omega)+jB_0(\omega)D_0(\omega)\Psi(z_0,\omega)-B_0(\omega)C_0(\omega)\Theta_{iv}^*(z_0,\omega)+A_0(\omega)D_0(\omega)\Theta_{iv}(z_0,\omega). \tag{21}$$

Taking the real and imaginary parts of Eq. (21), using the definitions (13) and (16), and introducing relation (4a), one finds

$$\Pi(z_1,\omega)=\Pi(z_0,\omega) \tag{22}$$

and

$$\Lambda(z_1,\omega)=A_0(\omega)C_0(\omega)\Phi(z_0,\omega)+B_0(\omega)D_0(\omega)\Psi(z_0,\omega)+[A_0(\omega)D_0(\omega)+B_0(\omega)C_0(\omega)]\Lambda(z_0,\omega). \quad (23)$$

Relation (22) is a consequence of the lossless character of the beam transducer which conserves kinetic power. The real part of the CPDS of the noise voltage and current is an invariant with respect to lossless transformations. It is not hard to see that another noise parameter exists which is invariant with respect to lossless transformations. Indeed, Eqs. (19), (20), and (23) show that the power density spectra, $\Phi(z,\omega)$, $\Psi(z,\omega)$, and $\Lambda(z,\omega)$, at the output of the lossless beam transducer are determined solely by the values of the same power density spectra at the input. Two processes with the same values of $\Phi(z_0,\omega)$, $\Psi(z_0\,\omega)$, and $\Lambda(z_0,\omega)$ transform in an identical manner. If in particular, one of the processes is a coherent process, then $\Phi(z_0,\omega)\times\Psi(z_0,\omega)-\Lambda^2(z_0,\omega)=\Pi^2(z_0,\omega)$. But, $\Pi(z_0,\omega)=\Pi(z_1,\omega)$, and correspondingly

$$\Phi(z_0\,\omega)\Psi(z_0,\omega)-\Lambda^2(z_0,\omega)=\Phi(z_1,\omega)\Psi(z_1\,\omega)-\Lambda^2(z_1,\omega). \quad (24)$$

Equation (24) does not hold for a coherent process only. It has been pointed out that the power density spectra, $\Phi(z,\omega)$, $\Psi(z\,\omega)$, and $\Lambda(z\,\omega)$ of two processes with the same input values $\Phi(z_0,\omega)$, $\Psi(z_0,\omega)$, and $\Lambda(z_0,\omega)$, transform in an identical way regardless of their value of $\Pi(z_0,\omega)$. Since Eq. (24) contains only the first three power density spectra, it must hold for any noise process. A somewhat lengthy mathematical proof of Eq. (24) can be given by the use of Eqs. (19), (20), (23), and (4a).

We choose a separate symbol for this second noise parameter which is invariant with respect to lossless beam transformations

$$S^2(z_0,\omega)=\Phi(z_0\,\omega)\Psi(z_0\,\omega)-\Lambda^2(z_0,\omega). \quad (25)$$

Equations (24) and (25) introduced into Eq. (17) give

$$Z_0^2\Psi_{\max}\Psi_{\min}=S^2(z_0,\omega). \quad (26)$$

Equation (26) states that the product $\Psi_{\max}\Psi_{\min}$ of the maximum and minimum of the standing wave of the current SPDS in the drift region is determined solely by the characteristic impedance of the beam in the drift region and the input conditions at the cross section z_0, independent of the lossless beam transducer used between z_0 and the drift region.†[8]

The investigation of this section leads to an interesting formulation of the following practical problem: Find the matrix of a lossless beam transducer between a cross section z_0 and a drift region of characteristic impedance Z_0 which transforms given input conditions, $\Phi(z_0,\omega)$, $\Psi(z_0,\omega)$, $\Lambda(z_0,\omega)$ and $\Pi(z_0,\omega)$, into a standing wave of the current SPDS with a minimum $\Psi_{\min}$ at a distance $\Delta\theta$ from the transducer output.

The problem can be reduced to a conventional problem of four-terminal network theory. The transducer in question causes the same standing-wave pattern of $\Psi(z,\omega)$ in the drift region when fed by a coherent process with the same input values, $\Phi(z_0,\omega)$, $\Psi(z_0,\omega)$, and $\Lambda(z_0,\omega)$, and with a real part of the CPDS, $\Pi_{eq}(z_0,\omega)=[\Phi(z_0,\omega)\Psi(z_0,\omega)-\Lambda^2(z_0,\omega)]^{\frac{1}{2}}$. Proportional to the real and reactive power of the coherent process, are $\Lambda(z_0,\omega)$ and $\Pi_{eq}(z_0,\omega)$. The coherent process has an impedance at the input

$$Z_{in}(\omega)=\left[\frac{\Phi(z_0,\omega)}{\Psi(z_0,\omega)}\right]^{\frac{1}{2}}e^{j\delta}$$

with $\delta=\tan^{-1}[\Lambda(z_0,\omega)/\Pi_{eq}(z_0,\omega)]$. The desired impedance at the output of the transducer, Z_{out}, can be computed from the transmission line analog of the beam.[9]

$$Z_{out}=\frac{jZ_0\tan(\Delta\theta)+R_{eq}}{Z_0+jR_{eq}\tan(\Delta\theta)}Z_0$$

where $R_{eq}=\Pi_{eq}(z_0,\omega)/[\Psi_{min}]^{\frac{1}{2}}$.

The appropriate beam transformer can now be found from the requirement that it transform the impedance Z_{out} into the impedance Z_{in}. The coefficients, $A_0(\omega)$, $jB_0(\omega)$, $-jC_0(\omega)$, and $D_0(\omega)$, of the transducer satisfy the reciprocity relation (4a). They can be determined by conventional methods of four-terminal network theory. The phase factor $e^{j\phi}$ which multiplies the matrix elements of a lossless nonreciprocal transducer [see Eqs. (4)] affects neither the impedance transformation (see reference 6), nor the transformation of the power density spectra and is therefore immaterial.

FIG. 3. Schematic of the velocity-jump amplifier: (a) standing wave of $\Psi(z_1\omega)$; (b) velocity-jump amplifier; and (c) time average potential distribution, $V_0(z)$.

† J. R. Pierce (see reference 8) proved relation (26) for the special case of a noise process with $\Pi(z_0,\omega)=0$.

[8] W. E. Danielson and J. R. Pierce, J. Appl. Phys. **25**, 1163 (1954).

[9] S. Bloom and R. W. Peter, RCA Rev. **15**, 252 (1954).

6. THE MINIMUM NOISE FIGURE OF A VELOCITY-JUMP AMPLIFIER AND A KLYSTRON

The noise in an electron beam determines the noise figure of an amplifier which employs the beam. The minimum noise figure of a velocity jump amplifier will be determined in this section. It will be assumed that the lossless beam transducer between the cross section z_0 beyond the potential minimum in the electron gun and the input of the amplifier is optimized. For the sake of simplicity a special model of a velocity-jump amplifier will be adopted (see Fig. 3). The signal is applied to the beam through a buncher cavity and picked up by a catcher cavity. The beam loading of the cavity gaps will be neglected under the assumption that the gap transit angle of an electron is small. The buncher gap is preceded by a drift region.

The noise figure of an amplifier at the frequency ω can be defined by

$$F=\frac{\text{noise power output per unit frequency}}{\text{noise power output per unit frequency in absence of beam noise}}. \quad (27)$$

The power output per unit frequency of the catcher cavity is proportional to $\Psi(z_3,\omega)$, where $\Psi(z_3,\omega)$ is the SPDS of the beam current at the position of the catcher cavity gap. The region between the two cavity gaps is a lossless beam transducer. At the exit of the buncher cavity gap at z_2 from Eq. (20), $\Psi(z_3,\omega)$ can be found in terms of the power density spectra;

$$\Psi(z_3,\omega)=C_0^2(\omega)\Phi(z_2,\omega)+D_0^2(\omega)\Psi(z_2,\omega)+2C_0(\omega)D_0(\omega)\Lambda(z_2,\omega). \quad (28)$$

The power density spectra, $\Phi(z_2,\omega)$, $\Psi(z_2,\omega)$, and $\Lambda(z_2,\omega)$, are caused by two independent, additive noise sources. On one hand, the thermal noise of the buncher cavity structure gives rise to a beam voltage modulation whose SPDS is equal to $kT\,\mathrm{Re}[Z(\omega)]/\pi$, where $Z(\omega)$ is the shunt impedance of the buncher cavity gap, k is Boltzmann's constant, and T is the temperature of the buncher cavity structure. On the other hand, the beam noise at the cross section z_1 appears at z_2 transformed by the action of the gap. A narrow gap with a shunt impedance $Z(\omega)$ is a beam transducer with the matrix

$$A(\omega)=1, \qquad B(\omega)=-Z(\omega), \qquad C(\omega)=0, \qquad D(\omega)=1, \quad (29)$$

The preceding discussion and the use of Eqs. (8), (9), (10), (13), and (29) lead to the results:

$$\Phi(z_2,\omega)=\Phi(z_1,\omega)+|Z(\omega)|^2\Psi(z_1,\omega)-2\,\mathrm{Re}[Z(\omega)]\Pi(z_1,\omega)-2\,\mathrm{Im}[Z(\omega)]\times\Lambda(z_1,\omega)+(1/\pi)kT\,\mathrm{Re}[Z(\omega)] \quad (30)$$

$$\Psi(z_2,\omega)=\Psi(z_1,\omega)$$

$$\Lambda(z_2,\omega)=\Lambda(z_1,\omega)-\mathrm{Im}[Z(\omega)]\Psi(z_1,\omega).$$

The cross section z_1 lies at the end of a drift region. It is advantageous to express $\Phi(z_1,\omega)$, $\Psi(z_1,\omega)$, and $\Lambda(z_1,\omega)$ in terms of the distance of z_1 from the minimum of the standing wave of the current SPDS, $\Psi(z,\omega)$, in the drift region. We denote this distance measured in terms of plasma wavelengths by θ. Making use of Eqs. (12), (14), (15), and (26), and noting that $\Phi_{\max}=Z_0^2\Psi_{\max}$, one finds

$$\Phi(z_1,\omega)=\cos^2\theta\frac{S^2(z_0,\omega)}{\Psi_{\min}}+Z_0^2\sin^2\theta\Psi_{\min}, \quad (31)$$

$$\Psi(z_1,\omega)=Y_0^2\sin^2\theta\frac{S^2(z_0,\omega)}{\Psi_{\min}}+\cos^2\theta\Psi_{\min}, \quad (32)$$

$$\Lambda(z_1,\omega)=\tfrac{1}{2}Z_0\left[\frac{Y_0^2S^2(z_0,\omega)}{\Psi_{\min}}-\Psi_{\min}\right]\sin 2\theta. \quad (33)$$

Equations (30), (31), (32), and (33) introduced into the definition of the noise figure (27) give

$$F=1+\frac{\pi}{kT\,\mathrm{Re}[Z(\omega)]}\left\{M\left[\frac{Y_0^2S^2(z_0,\omega)}{\Psi_{\min}}+\Psi_{\min}\right]+N(\theta)\left[\frac{Y_0^2S^2(z_0,\omega)}{\Psi_{\min}}-\Psi_{\min}\right]\right\}-\frac{2\pi}{kT}\Pi(z_1,\omega) \quad (34)$$

where

$$M=\tfrac{1}{2}\left\{Z_0^2+|Z(\omega)|^2+\frac{D_0^2(\omega)}{C_0^2(\omega)}-2\frac{D_0(\omega)}{C_0(\omega)}\,\mathrm{Im}[Z(\omega)]\right\}$$

and

$$N(\theta)=\tfrac{1}{2}\left\{Z_0^2-|Z(\omega)|^2-\frac{D_0^2(\omega)}{C_0^2(\omega)}+2\frac{D_0(\omega)}{C_0(\omega)}\,\mathrm{Im}[Z(\omega)]\right\}\times\cos 2\theta+Z_0\left\{\frac{D_0(\omega)}{C_0(\omega)}-\mathrm{Im}[Z(\omega)]\right\}\sin 2\theta.$$

We proceed to minimize the noise figure of expression (34) by an adjustment in the lossless beam transducer which precedes the drift region. A change in the beam transducer will change both $\Psi_{\min}$ and θ, but will leave $S^2(z_0,\omega)$ and $\Pi(z_1,\omega)=\Pi(z_0,\omega)$ invariant. Differentiating (34) with respect to $\Psi_{\min}$ we obtain

$$\Psi_{\min}=|Y_0|S(z_0,\omega)\left[\frac{M+N(\theta)}{M-N(\theta)}\right]^{\frac{1}{2}}. \quad (35)$$

Equation (35) introduced into Eq. (34) gives

$$F=1+\frac{2\pi}{kT\,\mathrm{Re}[Z(\omega)]}|Y_0|S(z_0,\omega)\times[M^2-N^2(\theta)]^{\frac{1}{2}}-\frac{2\pi}{kT}\Pi(z_0,\omega). \quad (36)$$

The noise figure expression (36) can be minimized with respect to θ. A minimum occurs when $N^2(\theta)$ is a maxi-

mum, N^2_{max}. It can be verified easily that

$$(M^2-N^2_{max})^{\frac{1}{2}}=Z_0\,\mathrm{Re}[Z(\omega)].$$

Thus one finds for the minimum noise figure of a velocity-jump amplifier

$$F_{min}=1+\frac{2\pi}{kT}[S(z_0,\omega)-\Pi(z_0,\omega)]. \tag{37}$$

The minimum noise figure (37) depends only upon the two noise parameters, $S(z_0,\omega)$ and $\Pi(z_0,\omega)$, and is independent of the particular lossless beam transducer used for amplification between the two cavities. A drift region is a lossless beam transducer. When the lossless beam transducer between the two cavity gaps of the velocity-jump amplifier in Fig. 3 is replaced by a drift region, the velocity-jump amplifier reduces to a simple two-cavity klystron. Accordingly, expression (37) also gives the minimum noise figure of a klystron amplifier with short cavity gaps obtainable by means of a lossless beam transducer acting upon the noise in the beam before it enters the amplifier.

7. THE MINIMUM OBTAINABLE NOISE FIGURE OF A TRAVELING-WAVE TUBE

The noise figure of a traveling-wave tube depends upon the amount of noise in the beam at the input z_1 of the slow wave structure of the tube. The four power density spectra $\Phi(z_1,\omega)$, $\Psi(z_1,\omega)$, $\Lambda(z_1,\omega)$, and $\Pi(z_1,\omega)$ determine the noise in the beam. An alternate description of the noise can be given in terms of another set of four parameters: The maximum Ψ_{max} of the SPDS of the noise current in the drift region preceding the slow wave structure, and its minimum Ψ_{min}; the distance of the minimum of the noise current SPDS from the input of the TWT measured in terms of plasma wavelengths, θ; and, finally, the value of the real part of the CPDS, $\Pi(z_1,\omega)$ constitutes such a set. If a lossless beam transducer is used between the reference cross section z_0 (see Fig. 1) and the drift region, the parameters $Z_0^2\Psi_{max}\Psi_{min}=S^2(z_0,\omega)$ and $\Pi(z_1,\omega)=\Pi(z_0,\omega)$ are fixed entirely by the noise process in the potential minimum region. In terms of the described parameters, the noise figure of a traveling-wave tube has the form (see Appendix IV)

$$F=1+\frac{4QC^2 4\pi V_0}{I_0kT}\left[\frac{Y_0^2S^2(z_0,\omega)}{\Psi_{min}}f(QC,b,d,\theta)+\Psi_{min}f(QC,b,d,\theta+\tfrac{1}{2}\pi)\right]-\frac{4\pi}{kT}g(QC,b,d)\Pi(z_0,\omega) \tag{38}$$

where the space-charge parameter QC, the velocity parameter b, and the loss parameter d are defined in reference 1. Further,

$$f(QC,b,d,\theta)=\left|(\delta_2+\delta_3)\cos\theta-\frac{1}{(4QC)^{\frac{1}{2}}}(\delta_2\delta_3-4QC)\sin\theta\right|^2 \tag{39}$$

and

$$g(QC,b,d)=\tfrac{1}{2}[j(\delta_2+\delta_3)(\delta_2\delta_3-4QC)^*-j(\delta_2+\delta_3)^*(\delta_2\delta_3-4QC)].$$

Pierce also defines δ_2 and δ_3 in reference 1; $f(QC,b,d,\theta)$ is a function of the parameters indicated in the parentheses. The function $f(QC,d,\theta)$ defined and plotted by Watkins[4] is obtained from $f(QC,b,d,\theta)$ by choosing the values of b which give maximum gain.

The parameters Ψ_{min} and θ in Eq. (38) can be adjusted by varying the lossless beam transducer between the reference cross section z_0 and the drift region. In particular, Ψ_{min} and θ can be so chosen that they minimize the noise figure of the traveling-wave tube. In order to find the minimization conditions, we rewrite the function $f(QC,b,d,\theta)$ in a different form. $f(QC,b,d,\theta)$ is a periodic function of θ, with a maximum f_{max} at some position θ_0 away from the input of the traveling-wave tube. Denote its minimum by f_{min}. Then,

$$f(QC,b,d,\theta)=\tfrac{1}{2}(f_{max}+f_{min})+\tfrac{1}{2}(f_{max}-f_{min})\cos 2(\theta-\theta_0).$$

Defining $\beta=(f_{max}-f_{min})/(f_{max}+f_{min})$ we can write the noise figure expression (38) in a more convenient form:

$$F=1+\frac{4QC^2 2\pi V_0}{I_0kT}(f_{max}+f_{min})\Psi_{min}\times\left\{[1-\beta\cos 2(\theta-\theta_0)]+\frac{Y_0^2S^2(z_0,\omega)}{\Psi_{min}{}^2}[1+\beta\cos 2(\theta-\theta_0)]\right\}-\frac{4\pi}{kT}g(QC,b,d)\Pi(z_0,\omega). \tag{40}$$

Minimization of Eq. (40) with respect to Ψ_{min} gives

$$\Psi_{min}=|Y_0|\,S(z_0,\omega)\left[\frac{1+\beta\cos 2(\theta-\theta_0)}{1-\beta\cos 2(\theta-\theta_0)}\right]^{\frac{1}{2}}. \tag{41}$$

Inserting Eq. (41) into expression (40) and minimizing with respect to θ, one finds

$$\theta=\tfrac{1}{2}\pi+\theta_0, \tag{42}$$

and correspondingly, from Eqs. (42) and (41)

$$\Psi_{min}=|Y_0|\,S_0(z_0,\omega)\left(\frac{f_{min}}{f_{max}}\right)^{\frac{1}{2}}. \tag{43}$$

Further, it can be shown that

$$g(QC,b,d)=(4QCf_{max}f_{min})^{\frac{1}{2}}. \tag{44}$$

The use of Eqs. (42), (43), and (44) in Eq. (40) results in an expression for the minimum noise figure obtainable by means of a lossless transducer;

$$F_{min}=1+\frac{4\pi}{kT}(4QCf_{max}f_{min})^{\frac{1}{2}}[S(z_0,\omega)-\Pi(z_0,\omega)]. \tag{45}$$

In the particular case in which the slow wave structure is lossless ($d=0$), it can be shown that the expression $(4QCf_{max}f_{min})$ has a minimum at the value of b at which maximum gain occurs

$4QCf_{max}f_{min}=\frac{1}{4}$, independent of the value of QC.

Thus, the minimum noise figure has the particularly simple form

$$F_{min}=1+\frac{2\pi}{kT}[S(z_0,\omega)-\Pi(z_0,\omega)]. \tag{46}$$

Expression (46) for the minimum obtainable noise figure of a traveling-wave tube with $d=0$ is identical with the expression (37) obtained for the minimum noise figure of a klystron and of a velocity-jump amplifier with narrow-gap cavities as in- and output.

In the hypothetical case in which the noise in the beam is a coherent process, with $S(z_0,\omega)=\Pi(z_0,\omega)$, the minimum obtainable noise figure is unity. It is possible to adjust the phases between the noise voltage and current modulations at the input of the traveling-wave tube to prevent the excitation of the growing wave. A general noise process has $S(z_0,\omega)>\Pi(z_0,\omega)$. The excitation of the growing wave cannot be prevented completely. The noise figure is greater than unity.

8. TRANSFORMATION BY MEANS OF LOSSY BEAM TRANSDUCERS

Thus far most of our attention has been devoted to the transformation of the noise process by means of lossless beam transducers. The question arises: Can the ultimate limit of the noise figures of a velocity-jump amplifier, a klystron, or a traveling-wave tube as expressed in Eqs. (37) and (45) be reduced if lossy beam transducers are used instead of lossless transducers, as heretofore assumed. The answer to this question is in the negative, as will be shown in this section.

The ultimate limit of the noise figure depends in part upon the particular type of amplifying structure used, in part upon the difference between the two noise parameters $S(z_0,\omega)$ and $\Pi(z_0,\omega)$. Their value is fixed entirely by the noise process in the potential minimum region.

A lossy transducer inserted between the cross section z_0 and the beginning of the drift region at z_1 transforms the parameters $S(z_0,\omega)$ and $\Pi(z_0,\omega)$ at its input to $S(z_1,\omega)$ and $\Pi(z_1,\omega)$ at its output. The proof that a lossy transducer does not improve the ultimate limit of the noise figure is completed when it is shown that

$$\left(\frac{S(z_1,\omega)-\Pi(z_1,\omega)}{S(z_0,\omega)-\Pi(z_0,\omega)}\right)_{min}\geqslant 1. \tag{47}$$

The most general lossy nonreciprocal four-terminal network can be represented by an equivalent circuit,[6] as shown in Fig. 4. The constant k of the ideal amplifier (or attenuator) is limited by the inequality

$$\frac{1}{2}\left(k+\frac{1}{k}\right)\leqslant(R_sG_p+1)^{\frac{1}{2}}. \tag{48}$$

In this paper the output quantities of a beam transducer are expressed by a matrix multiplication in terms of the input quantities as opposed to the convention used in reference 6. No major changes in the notation are necessary, however. The parameters S and Π are invariant with respect to a lossless transformation. Therefore, $S(z_0,\omega)$ and $\Pi(z_0,\omega)$ appear unchanged at the terminals "0" in Fig. 4, and correspondingly $S(z_1,\omega)$ and $\Pi(z_1,\omega)$ at the terminals "2".

Let us denote by Φ_1, Ψ_1, Λ_1, and Π_1 the various power density spectra of the noise process at the point "1" of the equivalent network in Fig. 4. With

$$S_1^2=\Phi_1\Psi_1-\Lambda_1^2 \tag{49}$$

one finds for $S^2(z_1,\omega)$, with the aid of the transformation matrix for the parallel conductance G_p in cascade with the ideal amplifier (or attenuator) and by means of Eqs. (8), (9), (10), (16), and (25),

$$S^2(z_1,\omega)=k^4(S_1^2+G_p^2\Phi_1^2-2G_p\Phi_1\Pi_1). \tag{50}$$

Similarly, one finds

$$\Pi(z_1,\omega)=k^2(\Pi_1-G_p\Phi_1). \tag{51}$$

In a similar way, $S(z_0,\omega)$ and $\Pi(z_0,\omega)$ can be expressed in terms of the power density spectra at the point "1";

$$S^2(z_0,\omega)=S_1^2+R_s^2\Psi_1^2+2R_s\Psi_1\Pi_1 \tag{52}$$

$$\Pi(z_0,\omega)=\Pi_1+R_s\Psi_1. \tag{53}$$

One finds for the ratio in Eq. (47)

$$\frac{S(z_1,\omega)-\Pi(z_1,\omega)}{S(z_0,\omega)-\Pi(z_0,\omega)}=k^2\frac{\left(1+\dfrac{G_p^2\Phi_1^2}{S_1^2}-\dfrac{2G_p\Phi_1}{S_1}\dfrac{\Pi_1}{S_1}\right)^{\frac{1}{2}}-\dfrac{\Pi_1}{S_1}+\dfrac{G_p\Phi_1}{S_1}}{\left(1+\dfrac{R_s^2\Psi_1^2}{S_1^2}+\dfrac{2R_s\Psi_1}{S_1^2}\dfrac{\Pi_1}{S_1}\right)^{\frac{1}{2}}-\dfrac{\Pi_1}{S_1}-\dfrac{R_s\Psi_1}{S_1}}. \tag{54}$$

The right-hand side of Eq. (54) can be minimized with respect to the parameters k, $G_p\Phi_1/S_1$, and $R_s\Psi_1/S_1$. According to the inequality (48), the smallest possible

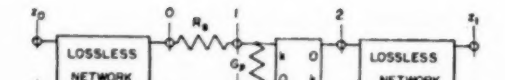


FIG. 4. Equivalent circuit of a lossy nonreciprocal beam transducer.

value of k is

$$k_{\min}=(R_sG_p+1)^{\frac{1}{2}}-(R_sG_p)^{\frac{1}{2}}. \tag{55}$$

The denominator of the right-hand side of Eq. (54) is a monotonically decreasing function of $R_s\Psi_1/S_1$. Its largest value occurs for the smallest possible value of $R_s\Psi_1/S_1$. Now,

$$\Psi_1=\frac{S_1^2+\Lambda_1^2}{\Phi_1}\geqslant\frac{S_1^2}{\Phi_1} \quad \text{or} \quad \frac{R_s\Psi_1}{S_1}\geqslant R_sG_p\frac{S_1}{G_p\Phi_1}. \tag{56}$$

The minimization of k and the maximization of the denominator result in

$$\frac{S(z_1,\omega)-\Pi(z_1,\omega)}{S(z_0,\omega)-\Pi(z_0,\omega)}\geqslant[(R_sG_p+1)^{\frac{1}{2}}-(R_sG_p)^{\frac{1}{2}}]^2 \times\frac{(1+x^2-2\delta x)^{\frac{1}{2}}-\delta+x}{\left[1+\left(\frac{R_sG_p}{x}\right)^2+2\delta\frac{R_sG_p}{x}\right]^{\frac{1}{2}}-\delta-\frac{R_sG_p}{x}}$$

with

$$x=(G_p\Phi_1)/S_1 \quad \text{and} \quad \delta=\Pi_1/S_1. \tag{57}$$

A further minimization with respect to x results in a fourth-order equation for x with only one acceptable root x_1

$$x_1=-\delta R_sG_p+[(\delta R_sG_p)^2+R_sG_p]^{\frac{1}{2}}. \tag{58}$$

By using the value of the root x_1 it is simple to show that the right-hand side of Eq. (57) is equal to unity. For this purpose R_sG_p can be expressed in terms of x_1 and δ,

$$R_sG_p=\frac{x_1^2}{1-2\delta x_1}. \tag{59}$$

The value for R_sG_p from Eq. (59) leads to the cancellation of the numerator and denominator in the inequality (57), independent of the values of x_1 and δ. The conclusion is that

$$\left(\frac{S(z_1,\omega)-\Pi(z_1,\omega)}{S(z_0,\omega)-\Pi(z_0,\omega)}\right)_{\min}=1. \tag{60}$$

The equality (60) completes the proof. It should be emphasized that no assumption has been made concerning the character of the noise. The values of S_1 and Π_1 and correspondingly $S(z_0,\omega)$ and $\Pi(z_0,\omega)$, have been left unspecified. It is also important to note that the minimum value of unity for the ratio in the equality (60) is achieved only if the lossless transducer satisfies the condition $(k+1/k)/2=(R_sG_p+1)^{\frac{1}{2}}$.

The investigation of this section has led to the generalization of the minimum noise figure expressions (37) and (45). It has been shown that these expressions apply to the minimization by means of lossless and lossy transducers alike. In the past, lossy beam transducers have been proposed for reduction of the noise figure of a traveling-wave tube.[10] A lossy beam transducer cannot lead to a noise figure lower than that of Eq. (46), subject to the assumptions of the one-dimensional, single-velocity theory.

9. CONCLUSIONS

The input conditions of the noise in a one-dimensional electron beam have been generalized to include the most general possible cases. The advantage of such a step was twofold. First, there is now no reliable high-frequency analysis of the potential minimum region in an electron gun that would give the noise input conditions at a cross section at and beyond which the single-velocity approximation is valid. If, and when, such a theory is developed, its results can be applied directly to the formalism of this paper. Second, the greater generality of the approach led to a remarkably simple formalism. A way was found by which a transformation of a general noise process by a lossless transducer can be reduced to a conventional problem of impedance transformation by a four-terminal network.

The minimum noise figures of a traveling-wave tube, and of specific models of a klystron and a velocity-jump amplifier attainable by means of the most general beam transducer of the four-terminal network type have been found in terms of two noise parameters $S(z_0,\omega)$ and $\Pi(z_0,\omega)$. These two parameters are determined entirely by the noise process in the electron gun in the region between the cathode and a reference cross section slightly beyond the potential minimum. The class of beam transducers of the four-terminal network type is wide, indeed. Beam expansions, velocity jumps, a noise-reducing prehelix terminated in passive loads, are all examples of passive beam transducers of the four-terminal network type. Any analysis of a specific four-terminal noise-reducing beam transducer on a one-dimensional basis cannot yield results that are more optimistic in predicting the noise figure of a klystron and a velocity-jump amplifier with narrow-gap cavities as their input and output, or a traveling-wave tube than the values given in Eqs. (37) and (46).

ACKNOWLEDGMENT

The author is indebted to Professor L. J. Chu for his inspiring guidance during the progress of this work. He wishes to thank Mr. L. D. Smullin and his associates

[10] H. Schnittger and D. Weber, Fernmelde tech. Z. **7**, 302 (1953).

of the Tube Laboratory of the Research Laboratory of Electronics, Massachusetts Institute of Technology, for many interesting discussions on the subject. He is grateful to Professor Y. W. Lee whose lucid teaching led the author to the present formulation of the problem.

APPENDIX I

The auto-correlation of the beam voltage modulation at the cross section z_2, $\phi(z_2,\tau)$, is given by

$$\phi(z_2,\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}V_1(z_2,t)V_1(z_2,t+\tau)dt.$$

Making use of Eq. (5a), one finds

$$\begin{aligned}\phi(z_2,\tau)=\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\Big[&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}a(\nu)a(\sigma)\\&V_1(z_1,t-\nu)V_1(z_1,t-\sigma+\tau)d\nu d\sigma\\&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}a(\nu)b(\sigma)V_1(z_1,t-\nu)\\&\times I_1(z_1,t-\sigma+\tau)d\nu d\sigma\\&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}b(\nu)a(\sigma)I_1(z_1,t-\nu)\\&\times V_1(z_1,t-\sigma+\tau)d\nu d\sigma\\&+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}b(\nu)b(\sigma)I_1(z_1,t-\nu)\\&\times I_1(z_1,t-\sigma+\tau)d\nu d\sigma\Big]dt.\end{aligned}\quad(\text{I.1})$$

We evaluate the third integral in Eq. (I.1) for illustration. An interchange of the order of integration and the limiting process[11] gives

$$\begin{aligned}&\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}b(\nu)a(\sigma)I_1(z_1,t-\nu)\\&\times V_1(z_1,t-\sigma+\tau)d\nu d\sigma dt\\&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}b(\nu)a(\sigma)\Big[\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}I_1(z_1,t-\nu)\\&\times V_1(z_1,t-\sigma+\tau)dt\Big]d\nu d\sigma.\end{aligned}\quad(\text{I.2})$$

The expression in brackets in Eq. (I.2) is the cross-correlation function between the voltage and the current in the beam at the cross section z_1;

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}I_1(z_1,t-\nu)V_1(z_1,t-\sigma+\tau)dt=\theta_{iv}(z_1,\tau+\nu-\sigma).$$

The third integral of Eq. (I.1) becomes

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}b(\nu)a(\sigma)\theta_{iv}(z_1,\tau+\nu-\sigma)d\nu d\sigma.$$

In a similar way all the other terms in Eq. (I.1) can be evaluated. The result is Eq. (7) of the text.

APPENDIX II

The Fourier transform of the auto-correlation function of the beam voltage modulation is the self-power density spectrum, $\Phi(z,\omega)$;

$$\Phi(z,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\phi(z,\tau)e^{-j\omega\tau}d\tau.$$

The Fourier transform of Eq. (7) can be carried out term by term. For illustration we carry out the transform of the last term of Eq. (7) in detail:

$$\begin{aligned}&\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}a(\sigma)b(\nu)\theta_{iv}(z_1,-\tau+\nu-\sigma)e^{-j\omega\tau}d\nu d\sigma d\tau\\&=\int_{-\infty}^{\infty}a(\sigma)e^{j\omega\sigma}d\sigma\int_{-\infty}^{\infty}b(\nu)e^{-j\omega\nu}d\nu\frac{1}{2\pi}\\&\times\int_{-\infty}^{\infty}\theta_{iv}(z_1,\mu)e^{j\omega\mu}d\mu\quad\text{with}\quad\mu=-\tau+\nu-\sigma.\end{aligned}$$

With the aid of Eqs. (6) and the definition of $\Theta_{iv}(z_1,\omega)$,

$$\Theta_{iv}(z_1,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\theta_{iv}(z_1,\tau)e^{-j\omega\tau}d\tau,$$

one finds for the last expression the simple form $A^*(\omega)B(\omega)\Theta_{iv}^*(z_1,\omega)$. The remaining integrals can be transformed in an analogous way. The result is Eq. (8) of the text.

APPENDIX III

The product of the maximum and the minimum of the standing wave of the current SPDS can be determined from Eq. (14) rewritten in a slightly different form:

$$\begin{aligned}\Psi(z_2,\omega)=&\tfrac{1}{2}\big[Y_0^2\Phi(z_1,\omega)+\Psi(z_1,\omega)\big]+\tfrac{1}{2}\big[\Psi(z_1,\omega)\\&-Y_0^2\Phi(z_1,\omega)\big]\cos 2\theta+Y_0\sin 2\theta\Lambda(z_1,\omega)\end{aligned}\quad(\text{III.1})$$

where $\theta=\omega_p(z_2-z_1)/u$.

[11] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* (John Wiley and Sons, Inc., New York, 1949).

The equation for the position of the maximum is obtained by differentiation of Eq. (III.1) with respect to θ,

$$[Y_0^2\Phi(z_1,\omega)-\Psi(z_1,\omega)]\sin 2\theta_{max} + 2Y_0\cos 2\theta_{max}\Lambda(z_1,\omega)=0. \quad \text{(III.2)}$$

The minimum, Ψ_{min}, lies at $\theta_{max}+\frac{1}{2}\pi$. We obtain from Eq. (III.1) the product $\Psi_{max}\Psi_{min}$.

$$\Psi_{max}\Psi_{min}=\tfrac{1}{4}[[Y_0^2\Phi(z_1,\omega)+\Psi(z_1,\omega)]^2 - \{[\Psi(z_1,\omega)-Y_0^2\Phi(z_1,\omega)]\cos 2\theta_{max} + 2Y_0\sin 2\theta_{max}\Lambda(z_1,\omega)\}^2]. \quad \text{(III.3)}$$

Equation (III.2) squared and subtracted from Eq. (III.3) finally gives

$$\Psi_{max}\Psi_{min}=Y_0^2[\Phi(z_1,\omega)\Psi(z_1,\omega)-\Lambda^2(z_1,\omega)].$$

APPENDIX IV. THE NOISE FIGURE OF A TRAVELING-WAVE TUBE

The definition (27) of the noise figure will be used. If the gain of the tube is high, the only wave of importance at the output is the exponentially growing wave. The axial field of the growing wave at the input is caused partly by the noise in the beam, partly by the circuit noise fed into the slow wave structure. The two sources of noise are uncorrelated.

The self-power density spectrum (SPDS) of the axial field of the growing wave, $\Xi_1(\omega)$, can be obtained by a method similar to the one used in Sec. 3:

$$\Xi_1(\omega)=|F(\omega)|^2\Phi(z_1,\omega)+|G(\omega)|^2\Psi(z_1,\omega) + 2\,\mathrm{Re}[F(\omega)G^*(\omega)]\Pi(z_1,\omega) + 2\,\mathrm{Im}[F^*(\omega)G(\omega)]\Lambda(z_1,\omega) + |H(\omega)|^2\Xi(\omega). \quad \text{(IV.1)}$$

The four power density spectra of the noise in the beam evaluated at the input of the traveling-wave tube, are $\Phi(z_1,\omega)$, $\Psi(z_1,\omega)$, $\Lambda(z_1,\omega)$, and $\Pi(z_1,\omega)$. The SPDS of the axial electric field caused by the circuit noise fed into the slow wave structure, is $\Xi(\omega)$;

$$F(\omega)=-\frac{\omega}{u}C\frac{\delta_2+\delta_3}{\left(1-\frac{\delta_2}{\delta_1}\right)\left(1-\frac{\delta_3}{\delta_1}\right)},$$

$$G(\omega)=-2j\frac{V_0\omega}{I_0 u}C^2\frac{\delta_2\delta_3-4QC}{\left(1-\frac{\delta_2}{\delta_1}\right)\left(1-\frac{\delta_3}{\delta_1}\right)},$$

$$H(\omega)=\frac{1}{\left(1-\frac{\delta_2}{\delta_1}\right)\left(1-\frac{\delta_3}{\delta_1}\right)}.$$

The beam is assumed to pass through a drift region before it enters the slow wave structure. The noise establishes a standing wave of the current SPDS in the drift region. The minimum of the standing wave, Ψ_{min}, is assumed to lie a distance θ measured in space-charge wavelengths in front of the input of the traveling-wave tube. By the use of Eqs. (31), (32), and (33), $\Xi_1(\omega)$ can be written in the form

$$\Xi_1(\omega)=\frac{S^2(z_2,\omega)}{\Psi_{min}}|F(\omega)\cos\theta-jY_0G(\omega)\sin\theta|^2 + Z_0^2\Psi_{min}|F(\omega)\sin\theta+jY_0G(\omega)\cos\theta|^2 + 2\mathrm{Re}[F(\omega)G^*(\omega)]\Pi(z_1,\omega)+|H(\omega)|^2\Xi(\omega). \quad \text{(IV.2)}$$

The SPDS of the axial field impressed by the circuit noise, $\Xi(\omega)$, can be related to the temperature T of the circuit, making use of the definition of C given in reference 1;

$$C^3=\frac{|E^2|}{\left(\frac{\omega}{u}\right)^2 P}\frac{I_0}{4V_0}. \quad \text{(IV.3)}$$

The mean-square fluctuations per unit frequency of a quantity are related to its SPDS by a factor of 4π; $|E^2|/\Delta f=4\pi\Xi(\omega)$. Assuming that the slow wave structure is matched to the input,

$$P=kT\Delta f. \quad \text{(IV.4)}$$

Combining Eqs. (IV.3) and (IV.4), one finds

$$\Xi(\omega)=\frac{4V_0}{I_0}\left(\frac{\omega}{u}\right)^2 C^3\frac{kT}{4\pi}.$$

The power output of the traveling-wave tube is proportional to $\Xi_1(\omega)$. In the absence of the beam noise one would have

$$\Xi_1(\omega)=|H(\omega)|^2\Xi(\omega).$$

The noise figure of the traveling-wave tube is, according to the definition (27),

$$F=\frac{\Xi_1(\omega)}{|H(\omega)|^2\Xi(\omega)}=1+\frac{I_0kT}{4QC^2 4\pi V_0}\left[\frac{\Psi_{min}}{Y_0^2S^2(z_0,\omega)} \times f(QC,b,d,\theta)+\Psi_{min}f(QC,b,d,\theta+\tfrac{1}{2}\pi)\right] - \frac{4\pi}{kT}g(QC,b,d)\Pi(z_1,\omega) \quad \text{(IV.5)}$$

where

$$f(QC,b,d,\theta)=\frac{1}{C^2\left(\frac{\omega}{u}\right)^2}\left|\frac{F(\omega)}{H(\omega)}\cos\theta-jY_0\frac{G(\omega)}{H(\omega)}\sin\theta\right|^2$$

$$=\left|(\delta_2+\delta_3)\cos\theta-\frac{1}{(4QC)^{\frac{1}{2}}}(\delta_2\delta_3-4QC)\sin\theta\right|^2$$

using the fact that

$$Y_0=-\frac{I_0}{2V_0}(4QC^3)^{-\frac{1}{2}}.$$

Further,

$$g(QC,b,d)=\frac{-I_0}{2V_0\left(\frac{\omega}{u}\right)^2C^3}\mathrm{Re}\left[\frac{F(\omega)G^*(\omega)}{|H(\omega)|^2}\right]$$

$$=\tfrac{1}{2}[j(\delta_2+\delta_3)(\delta_2\delta_3-4QC)^*-j(\delta_2+\delta_3)^*(\delta_2\delta_3-4QC)].$$

It is assumed that the beam transducer in front of the traveling-wave tube is lossless. Therefore, $\Pi(z_1,\omega)=\Pi(z_0,\omega)$ and Eq. (IV.5) is found to be identical with Eq. (38) presented in the text.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Arcing at Electrical Contacts on Closure. Part VI. The Anode Mechanism of Extremely Short Arcs

W. S. Boyle and L. H. Germer
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

(Received July 19, 1954)

The erosion at the anode of a short arc has been studied as a function of the arc duration and the magnitude of the arc current. With the current maintained at a constant value the anode pit grows continuously in diameter up to a critical value of the order of 10 microns; when this size is reached, a new arc is initiated and a new pit starts to develop. The process may be repeated. Consideration of the heat loss from the anode shows that for a constant arc power the pressure of metal vapor between the electrodes is very sensitive to the area covered by the arc. An order of magnitude estimate of the minimum metal vapor pressure required to maintain the arc (~100 atmospheres) yields a value of the maximum pit diameter in good agreement with the experimental observations. It is evident that the steady-state short arc does not exist, and quantities such as the metal transfer, ion to electron current ratio, and pressure in the arc are dependent on the time since the initiation of the arc.

THIS paper reports an extension of the work of Germer and Haworth on the erosion of electrical contacts.[1] In their paper Germer and Haworth describe a series of experiments where a pair of electrical contacts in parallel with a small capacitor charged to some 50 volts are slowly brought together. It is shown that at a distance of the order of a few thousand angstroms an electrical discharge takes place which discharges the capacitor in a time approximately equal to the half period of the circuit. This discharge, thought to be a field emission type of arc,[2] produces a hemispherical pit on the anode surface and covers a small region of the cathode with an irregular deposit of metal which has been spattered from the anode. In the experiments to be described here the discharge was produced in a different manner. The electrodes were set at some particular separation still in the range of a few thousand angstroms and then a single square wave current pulse was driven across the gap between the contacts by a pulse generator with effectively infinite internal impedance. The electrode separation, and the size and duration of the arc current may be independently controlled, and, since the arc current is constant during a single discharge, conditions are simplified to a certain extent.

These experiments show that for a given current the arc grows continually in size up to some critical value and is then suddenly extinguished, a new arc is immediately formed and the process is repeated. The measured anode pit diameters yield values of the current density in the going arc (a function of the arc time), and numerical calculations of the heat transfer processes from the anode give a rough estimate of the pressure in the arc. The maximum arc diameter is determined by the minimum permissible pressure of metal vapor in the arc.

THERMAL PROCESSES AT THE ANODE OF A "SHORT ARC"

For the mechanism of the short arc the model described by Kisliuk will be used. It is essentially a field-emission type of arc where a positive ion space charge in front of the cathode produces a sufficiently large field to induce field emission. This type of arc is particularly efficient in its use of positive ions and in

[1] L. H. Germer and F. E. Haworth, J. Appl. Phys. **20**, 1085 (1949).
[2] P. P. Kisliuk, J. Appl. Phys. **25**, 897 (1954).

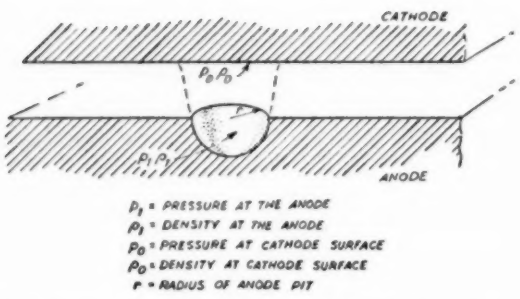

CATHODE
ANODE
p_1 = PRESSURE AT THE ANODE
ρ_1 = DENSITY AT THE ANODE
p_0 = PRESSURE AT CATHODE SURFACE
ρ_0 = DENSITY AT CATHODE SURFACE
r = RADIUS OF ANODE PIT

FIG. 1. Cross section of the electrodes showing an anode pit.

general the ratio of electron to ion current is large. A large fraction of the electrons cross between the electrodes without losing much energy by collisions and to a first approximation the power delivered to the anode will be just Iv, where v is the arc voltage ($\cong$15 volts) and I is the arc current.[3] Because of the small electron energies the collision cross section for ion production is small and this coupled with the small electrode separation requires a high pressure in the arc. This high pressure is produced by evaporating metal from the anode and this pressure will now be calculated as a function of power input and arc diameter at the anode. Immediately below are calculations of the heat lost from the anode spot by conduction and by evaporation. From a preliminary calculation it is known that the heat lost by radiation is negligible.

It is observed experimentally that the anode pits of single arcs are roughly hemispherical in shape. For this geometry the thermal resistance R from the surface of the pit into the body of the anode metal is given by $R=1/2\pi\kappa r$, where r is the radius of the pit and the thermal conductivity κ is assumed to be temperature independent. To show that the thermal conduction loss is not very sensitive to the shape of the pit, one may consider the thermal resistance of a flat circular plane on the surface of the anode. This is given by $R=1/4\kappa r$. Even this extreme case yields a value only some 50 percent larger than the spherical geometry. When the surface of the metal in the anode pit is at the boiling point T_B the rate of heat loss by conduction P_c will be

$$P_c=T_B 2\pi\kappa r \text{ for a spherical pit,} \qquad (1)$$

or

$$P_c=T_B 4\kappa r' \text{ for a flat pit.}$$

The simple model for calculating the rate of evaporation of metal from the anode pit is shown in Fig. 1. The pressure p_1 of the metal vapor at the orifice of the metal pit will be greater than the pressure p_0 near the cathode, since most atoms will condense on the relatively cool cathode surface. We have considered these two pressures in order to make use of the fact that if $p_1>2p_0$ the gas flow across the diametral plane of the hemisphere will be independent of p_0 and will be given by $q=0.9C_1\rho_1\pi r^2$ g/sec,[4] where C_1 is the velocity of sound at the orifice and ρ_1 is the density of the vapor in this region. If we assume the vapor obeys the perfect gas law $C_1=(\gamma p_1/\rho_1)^{\frac{1}{2}}$, where γ=ratio of the specific heats, $\rho_1=Mp_1/RT_B$, R=gas constant$=8.3\times10^7$ ergs/°C mol, and M=molecular weight. Hence $q=(3.5)r^2\rho_1\times(M/RT_B)^{\frac{1}{2}}$ g/sec. In order to convert into a heat current, the mass current q must be multiplied by a factor η which is the energy required to raise one gram of the metal to the boiling point and evaporate it. The heat loss by evaporation will be

$$P_e=3.5\eta r^2\rho_1(M/RT_B)^{\frac{1}{2}}. \qquad (2)$$

For gold, Eqs. (1) and (2) become $P_c=4.1\times10^4 r$ watts and $P_e=2.5\times10^5 r^2 p_1$ watts. Equating the sum of these two terms to the total arc power Iv yields a family of curves of p *vs* r for various values of the arc power. These curves are presented in Fig. 2. Because of the steepness of the p *vs* r curves, any crude estimate of the expected minimum pressure yields a fairly accurate value of the pit size, and from Fig. 2 we see the pit radii must lie in the region of 5 to 10×10^{-4} cm.

Thermal Equilibrium and Relaxation

In calculating the thermal loss by conduction it was tacitly assumed that a steady-state condition had been

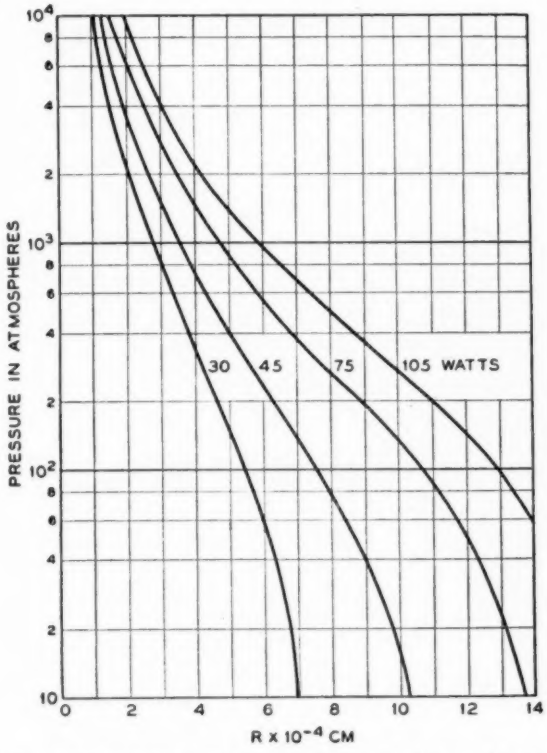


FIG. 2. Calculated metal vapor pressure at the anode as a function of the pit radius.

[3] Oscillograms showing a short arc potential difference, which does not change observably during the life of the arc, have been reproduced as Figs. 1B and 1C in J. Appl. Phys. **22**, 955 (1951). No comparable record was easily obtainable in the present experiments, but we believe that the arc voltage was substantially constant during each arc.

[4] H. Lamb, *Hydrodynamics* (Cambridge University Press, Cambridge, England, 1932), sixth edition, p. 27.

reached; if other than steady-state conditions prevail, the thermal loss will be larger than Eq. (1) indicates. This restriction does not apply to the loss by evaporation as given by Eq. (2).

The thermal relaxation time can be evaluated for the case of spherical symmetry and this will give also an indication of the time for the steady-state condition to be reached. An approximate solution of the heat flow problem for this geometry is given by $u=rT_B/(\pi Kt)^{\frac{1}{2}}$, where T_B is the initial temperature of the surface of the pit, K is the diffusivity, and t is the elapsed time. For the experimental conditions to be described, $r \doteqdot 5\times10^{-4}$ cm and $K \doteqdot 1$, so that, for $u=(\frac{1}{2})T_B$, $t \doteqdot 2.5\times10^{-7}$ sec.

The current pulses used in these experiments ran from 0.3 to 3.0 microseconds, so that the steady state may not be attained and the conduction loss as given by Eq. (1) is somewhat on the small side. One notes also from this calculation that the metal will remain molten in the pit while the arc is extinguished during a current pulse since the time it remains extinguished is certainly no more than 10^{-8} second, this time being calculated from the time it takes the pulse current to charge up the local capacity of the contacts to the breakdown voltage (40 $\mu\mu$f to 300 volts at 2 amp).

EXPERIMENTAL RESULTS

All the results presented here were obtained with gold electrodes. The cathode was a length of 0.75-mm gold wire with a rounded end, the anode a 3-mm cylinder with a carefully polished base. By mounting the anode on a table which could be moved in two dimensions with micrometer screws, it was possible to lay down an array of some 100 pits so that they could be easily located and identified when examined under a microscope. The experimental observations can be divided into two categories, transitions from single to multiple pits, and pit size.

FIG. 3. Double pit in the anode. Pulse current 3 amp, 1.0 μsec, $\times$2000.

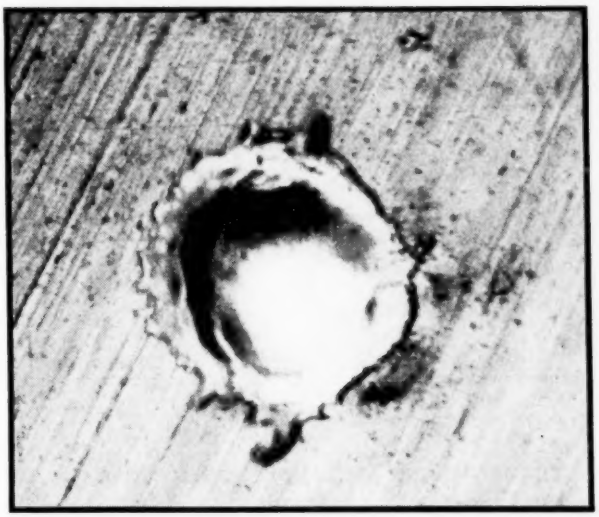


FIG. 4. Single pit in the anode. Pulse current 3 amp, 0.3 μsec, $\times$2000.

Transition from Single to Multiple Pits

The distinction between a single and a multiple pit is quite clear as shown in Figs. 3 and 4. Apart from the appearance of the separated pits there is a much larger amount of splashed out metal when a second arc is struck. Presumably during an arc the molten metal is swept out to the edge as soon as it is formed leaving a smooth pit, whereas, if a second arc is struck, during its initial stages it finds a pool of molten metal in the pit and because of the large pressure gradients molten metal is blasted out a great distance from the rim of the pit.

Reference to Table I shows a transition from single to multiple pits as the pulse length is increased for a given pulse current. The pit diameter is evidently growing with time and since the power input to the pit remains constant, the pressure of the evaporated metal vapor is continuously decreasing. (See Fig. 2). At some critical pressure insufficient ions are produced to maintain the arc, and it is extinguished. The extinguishing process is a regenerative one and the figures presented in Table I are probably very sensitive to the internal impedance of the pulse generator. This was qualitatively verified in an experiment using a length of transmission line as the pulse source. This had an internal impedance of 50 ohms which would be several orders of magnitude smaller than the internal impedance of the electronic pulse generator. In this case multiple pits occurred at pulse times which were smaller by a factor of 10 than those shown in Table I.

The fact that multiple pits appear at shorter pulse times for the smaller electrode separation is in further agreement with the concept of a continuously decreasing

TABLE I. Pit diameter in cm×10⁻⁴ for various current pulses.

Pulse current, amperes	0.3 μsec		0.7 μsec		1.0 μsec		2 μsec		3 μsec	
Electrode separation 30 000 A										
2	4.8,6.4	6.4,6.4	a a	a b	a a	a a	a a	a a	a a	a a
3	11,11,	8.0,9.5	16,13,	16,17	a a	a a	a 19	a 16	a a	a a
5	13,11,	16,16	a a	a a	a a	b a	a a	a a	a a	a a
7	a a	16, a	a a	a a	a a	a b	a a	a a	a a	a a
Electrode separation 15 000 A										
2	a a	6.5,7.0	8.0	8.0	a a	a a	a a	11, a	a a	a a
3	a a	13,11	a b	a b	a b	a b	11, b	b a	a a	a b
5	a b	16, b	b b	b b	b b	b b	a a	b a	b a	a b
7	16 b	b b	b a	a b	b a	a b	b a	b b	b b	b b

a Multiple pit.
b Electrodes welded.

pressure in the arc. At the smaller separation a larger pressure is required to maintain the steady arc and the critical value of the pressure is reached sooner.

Maximum Pit Size and Minimum Current Density

From the expected value of electron collision cross section one can estimate that the minimum vapor density in a short arc lies between normal atmospheric density and a value perhaps twenty times as large ($p=20$ atmos to $p=400$ atmos). Because in the curves of Fig. 2 the pressure is such a sensitive function of pit size, the maximum size can then be estimated quite accurately from these curves. The measured values of the maximum size of single pits taken from Table I, and written down in the third column of Table II, are in pretty good agreement with these theoretical estimates. Even if we were to assume an infinitely large collision cross section, the upper limit to the possible pit size obtained from such an assumption would differ from the experimental values by no more than about 50 percent.

Experimental values of the minimum current density in a short arc, obtained by dividing the currents in the first column of Table II by maximum pit areas, are written down in the last column of the table. As shown by Kisliuk[2] the current density can yield an estimate of the ratio of positive ion to electron current in a field-emission type of arc. The minimum current densities presented here are considerably smaller than those used in his paper and correspondingly the ratio of ion to electron current must be larger. This ratio is continually increasing throughout the history of a particular arc and the *maximum* ratio as calculated from the last column of Table II indicates that up to 15 percent of the arc current may be carried by ions.

DISCUSSION

A rather simple picture has been developed for some of the characteristics of a short arc. The observations have been made upon arcs between gold electrodes in which the current is strictly constant, but the conclusions no doubt apply also to more usual cases in which the current varies with time and to electrodes of other metals. Numerical values must be different for different metals.

At the beginning of a short arc the cross section is small, and the current density and the metal vapor pressure are high. The arc grows continuously in cross section until the current density has declined to a critical value, which is about 2 or 3×10^6 amp/cm² for gold. The arc must then be extinguished because of lack of sufficient vaporization from the anode, and if the current is maintained the arc will at once be struck again and the process of growth repeated. These conclusions are based upon calculations of the loss of heat from the anode, and they have been verified quantitatively by direct observations of anode pit size. During the initial stages of each arc the current is carried almost entirely by electrons, but during the life of the arc, the ion-electron ratio increases continuously to a maximum value of 0.15 when the arc is extinguished. The ion-electron ratio used by Kisliuk is a sort of average for the entire arc time, and is considerably smaller than this maximum.

TABLE II.

Pulse current Amperes	Electrode separation Angstroms	Maximum pit diameter cm×10⁻⁴	Minimum current density in a short arc Amperes/cm²
2	30 000	6.4	6×10^6
3	30 000	17	1.3×10^6
5	30 000	16	2.5×10^6
7	30 000	16	3.5×10^6
2	15 000	8	4.0×10^6
3	15 000	13	2.3×10^6
5	15 000	16	2.5×10^6
7	15 000	16	3.5×10^6

At the beginning of the arc the high-vapor pressure implies frequent collisions of electrons with metal atoms. Yet most of the energy of the arc is dissipated upon the anode and therefore most of the electron atom collisions must be elastic. When the arc is near its end the percentage of collisions resulting in ionization must be very greatly increased, but still most of the energy reaches the anode because of the reduced collision frequency. The great increase in the probability of ionization at a collision during the later stages of the arc is presumably caused by the increase in the arc voltage above the minimum which will just cause ionization. Since the difference between the arc voltage and the threshold voltage for ionization is small or about one volt (Kisliuk[2]), this effect may be very difficult to observe experimentally.

Multiple Coincidence Magnetic Storage Systems

ROBERT C. MINNICK AND ROBERT L. ASHENHURST
The Computation Laboratory of Harvard University, Cambridge, Massachusetts
(Received September 23, 1954)

In existing magnetic matrix storage systems a given location is selected by applying to two intersecting wires a current equivalent to one-half of the selecting field. This situation is generalized so that any core is selected by energizing p wires each with a current equivalent to $1/p$ of the selecting field. The advantages gained are in the correspondingly smaller fields applied to the nonselected cores, or alternatively, in the faster switching times obtainable by applying a total field greater than the coercive force to the selected core. To obtain the increased selection ratio $p-2$, additional sets of wires must be introduced into the matrix. Methods for accomplishing this are developed with the aid of a system analogous to plane analytic geometry for a finite set of points. Utilizing the increased selection ratio new fabrication techniques may be employed. Specifically, a storage matrix is illustrated in which the cores are toroids etched from a continuous sheet of magnetic material.

IN the last decade, increasing emphasis has been placed on the use of the hysteretic properties of magnetic material for the storage of information. This has been largely because of research in the field of digital computing machinery. Early developments in this direction were concerned with magnetic drum systems,[1] which are in extensive use at the present time. Such systems are characterized by capacities of the order of one million bits and random access times measured in milliseconds. The desire for somewhat shorter access times in systems of moderate capacity led to the idea of the static magnetic delay line[2–5] which differs from the drum in that it uses discrete magnetic cores rather than a continuous magnetic medium to store its information. Besides performing a storage function, the static delay line can also be used as a logical element, for example, as a serial shifting register or as a transducer between serial and parallel components or components operating at different frequencies.[6] Its principles of operation have been extended to the synthesis of switching devices of a very general character.[7]

A system of static delay lines, although capable of operating at frequencies in the vicinity of 100 kc, is still basically serial in nature. Its random access time may be reduced by decreasing the length of the lines and increasing their number, but there is a practical limitation because each line is controlled by a separate selection lead.

The idea of employing multiple coincidence to obtain random access in a core system without unduly expensive selection circuitry was put forth independently by J. W. Forrester[8] and J. A. Rajchman.[9,10] As in the case of the delay line, the property of the magnetic material employed for detecting its stored information is the difference in its response to signals of assigned polarities, depending on its remanent state. In addition, the coincidence systems utilize the fact that the response to signals of different amplitudes is highly nonlinear.

Most systems of this type evolved to date consist basically of a square matrix of n^2 cores threaded by a grid of n horizontal and n vertical selection wires. Each core is kept in one of its two remanent states, which may be arbitrarily designated by the binary digits 0 and 1. Reading information into a given core is accomplished by impressing currents on the horizontal and vertical wires threading it, in such a manner that the resulting fields reinforce. Let the field associated with each wire be $\frac{1}{2}H_1$, where H_1 exceeds the coercive force of the magnetic cores used. Then the procedure described will drive the core to state 0 or to state 1, depending on the current directions. The other cores which are threaded by either of the selection wires will be subjected to a field of magnitude $\frac{1}{2}H_1$; in order to insure that their information is not destroyed it is therefore necessary that $\frac{1}{2}H_1$ be considerably less than the coercive force, the exact amount depending on the squareness of the hysteresis loop involved. Reading out is accomplished by applying a sequence of pulses to a selected core, and sensing the voltage induced on an output winding which threads all the cores. Since this readout is destructive, regeneration circuitry must be provided to switch the core back to its original state, if necessary.

Two difficulties are commonly encountered in this type of operation. One is that the accumulation of small signals from the disturbed cores on the output wire tends to mask the desired output voltage. This effect can be lessened by threading the output wire with alternately aiding and opposing polarities and by using

[1] Staff of the Computation Laboratory, *Description of a Magnetic Drum Calculator*, Annals of the Computation Laboratory (Harvard University Press, Cambridge, Massachusetts, 1952), Vol. XXIV.
[2] Aiken, Wang, and Woo, "Static magnetic recording," *Progress Report No. 1, Investigation for the Design of Digital Calculating Machinery*, Harvard Computation Laboratory (August 10, 1948).
[3] Kincaid, Alden, and Hanna, Electronics **24**, No. 1, 108 (1951).
[4] A. Wang and W. D. Woo, J. Appl. Phys. **21**, 49 (1950).
[5] Kodis, Ruhman, and Woo, Convention Record, Proc. Inst. Radio Engrs. **1**, 38 (1953).
[6] Staff of the Computation Laboratory, *Description of the Mark IV Calculator*, Annals of the Computation Laboratory, Vol. XXVIII, (to be published).
[7] R. C. Minnick, J. Appl. Phys. **25**, 479 (April 1954).
[8] J. W. Forrester, J. Appl. Phys. **22**, 44 (1951).
[9] J. A. Rajchman, RCA Rev. **13**, 183 (1952).
[10] J. A. Rajchman, Proc. Inst. Radio Engrs. **41**, 1407 (1953).

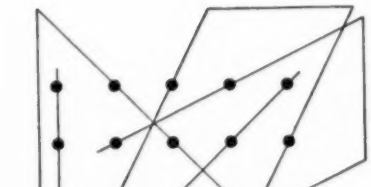


FIG. 1. Matrix with a typical member from each of six families of selection wires.

a double pulsing scheme and integrating the output signal as described by Rajchman.[10] The effect is nevertheless still present to some degree, and limits the size of the matrix obtainable with a given type core. The other difficulty is that disturbing fields applied to the nonselected cores may alter their states slightly, and if the hysteresis loop of the magnetic material is not sufficiently square successive disturbances may cause a core to lose its information altogether.

Both of these difficulties can apparently be minimized to any desired extent by subjecting the unselected cores to a field which is a smaller fraction of the selecting field H_1. It is convenient to introduce the term *selection ratio* p to denote the ratio of the magnitudes of the selecting and disturbing fields. In the square matrix described, $p=2$. It is possible to design systems in which the disturbing signal is carried on only one wire but for which the selection ratio can be made arbitrarily great if n is sufficiently large. A set of such systems will now be described.

Consider Fig. 1, representing twenty-five cores in a square array. Six wires are shown, all intersecting the core at the lower left. It should be noted that each other core is intersected by one and only one of these wires, and that each wire intersects five cores. If all six wires are energized, the lower left core is selected and all other cores in the matrix are disturbed. The ratio of the selecting to disturbing fields is $p=6$.

Two of these six wires are the usual horizontal and vertical wires while the other four are not found in the normal matrix. It will be shown that each of the six is a typical member of a family of wires, which can be introduced in such a way that each core is threaded by one and only one wire from each family. Furthermore, no two of the six wires associated with any core thread any other core. Therefore a given core can be selected with a ratio $p=6$ by impressing a current $1/6\ I_1$ on each of the six wires, one from each family, which thread the core.

To study this situation it is convenient to develop a special analytic geometry of $n\times n$ arrays. This will first be introduced as an abstract system, and then it will be shown how the results are applicable to the core matrices under discussion. No attempt will be made here to give a rigorous treatment; the discussion should be regarded merely as an outline of results which have been proved elsewhere.[11] In what follows, all numbers and arithmetic operations are to be taken modulo n.

Consider an $n\times n$ square array of elements, and let the rows and columns of the array each be numbered $0, 1, \cdots, n-1$ starting from the lower left. Then each element has associated with it two coordinates, x and y, corresponding to its column and row, respectively. Such an element, specified by its coordinates (x,y), will be called a *point*. In particular, the point (0,0) will be referred to as the *origin*. Given any point (x,y) and a number pair (a,b), where a and b are both relatively prime to n, the set of points $(x+ka, y+kb)$ $k=0, 1, \cdots, n-1$ are said to form a *line*. The fact that a and b are relatively prime to n insures that a line contains exactly n points, one in each row and column of the array. Note that the set of points comprising a row or a column do not correspond to lines under this definition. The pair (a,b) will be called a *generating pair* for the line; obviously the same line can be generated from several different sets of points and generating pairs.

The solution λ of the congruence $a\lambda\equiv b$ is defined as the *slope* of the line generated by (a,b). This solution always exists by virtue of the fact that a and n are relatively prime, and it can be shown that any line has a unique slope λ associated with it. A line through the origin with slope λ includes the point $(1,\lambda)$. Any line L can be uniquely specified by a point (x,y) and a slope λ. Given a point (x',y') not on L; the line L' through (x',y') with slope λ is said to be *parallel* to L, and has no point in common with it. Lines parallel to a given line are parallel to each other, and to every line there are exactly $n-1$ parallels, forming a family of n parallel lines which together include all of the n^2 points.

For the present purpose it is desired to find a set of lines through a given point which have no other point in common; such a set will be called *compatible*. It can be shown that if a set of lines through the origin is compatible, then the parallels to these through any other point will also be compatible. Furthermore, it can be demonstrated that the largest compatible set through the origin consists of n_0-1 lines, where n_0 is the least prime factor of n. The set of n_0-1 lines can usually be chosen in several ways; the condition which must be satisfied is that the set of slopes all be distinct modulo n_i, where n_i is any prime factor of n. Thus the set of lines with slopes $1, 2, \cdots, n_0-1$ is always a compatible set.

If n is prime, $n_0=n$ and the largest compatible set contains $n-1$ wires. The four skew wires shown in Fig. 1 correspond to the largest compatible set of lines

[11] R. L. Ashenhurst and R. C. Minnick, "Multiple coincidence magnetic storage systems," Harvard Computation Laboratory, *Bell Laboratories' Report No. 4, Theory of Switching* (July, 1953).

through the origin for $n=5$. By the foregoing results, if the wires corresponding to the parallels to these are introduced into the matrix, with the remaining horizontal and vertical wires, there will be six wires threading every core, no two of which thread any other core. Thus a selection ratio $p=6$ is attained. In general, the maximum selection ratio is $p=n_0+1$, representing wires corresponding to the largest set of compatible lines plus the horizontal and vertical lines. It is easily seen that an upper bound for the number of wires is given by $n+1$, since this reflects the case where one wire threads each core in the matrix. This upper bound is only reached for n prime by the present system. Using wires which do not correspond to lines in the present sense, it is possible to attain this upper bound in certain other cases, but the wiring scheme is more complicated.

The easiest way to determine the cores threaded by a given wire is by counting, using a convenient generating pair (a,b). For example, the wire corresponding to

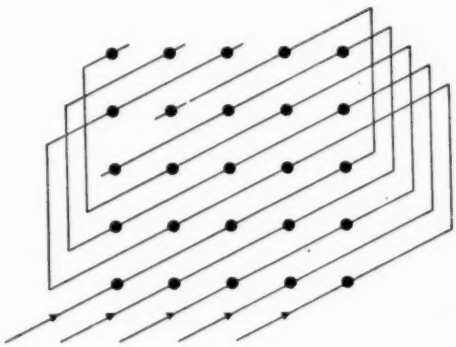


FIG. 2. Matrix with all members of the (2,1) family of selection wires.

the line of slope $\lambda=3$ is generated by the pair (2,1) in Fig. 1. Starting at the origin, a count of two horizontally and one vertically is made; on reaching the right-hand edge of the matrix the count is continued as if the first column were adjacent to the fifth. The corresponding points generated are therefore (0,0), (2,1), (4,2), (1,3), (3,4), and it can be verified that the wire in question threads the corresponding cores. This wire and its parallels are shown separately in Fig. 2.

Obviously, if n is even, the greatest selection ratio possible by this system is $p=3$; if n is odd but divisible by three the greatest ratio is $p=4$, and so on. It is desirable to pick sets of wires such that the distance between successive cores on a selection wire is small, so that useful generating pairs are (1,1), (1,−1), (1,2), (2,1), (−1,2), (2,−1), (2,3), etc.

The fact that the selection ratio can be increased in this manner allows a reconsideration of the types of magnetic material used for storage, since the demands of uniformity and squareness on the material are reduced. This introduces the possibility of using, instead of discrete cores, a continuous sheet of magnetic

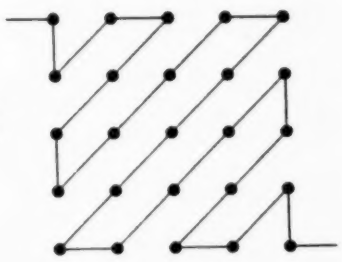


FIG. 3. Standard method for threading the output wire.

material pierced with holes through which wires are threaded. This, however, produces a new constraint on the sets of wires which can be threaded through the matrix. In the normal manner of threading the horizontal and vertical wires, selecting current for successive holes on a wire is directed alternately into and out of the face of the sheet. This essentially assigns a polarity to each hole and any additional wires must be threaded so as to be consistent with these polarities. It is easily verified that if a generating pair is such that $a+b$ is odd, then the corresponding wires will obey this constraint. Clearly wires can be put through holes at the edges of the matrix with either polarity.

For all such systems the set of wires corresponding to the pair (1,1) does not obey this restriction; hence if n is even, it is impossible to thread a third set of wires so that the criterion is obeyed. If n is odd, however, it is always possible to introduce two additional sets of wires, and the sets (1,2) and (1,−2) will always be acceptable. It also can be shown that if the least prime factor of n is five or more, then the sets (1,2) and (2,1) may always be used.

In designing a system with the aid of this technique, the number of extra sets of wires required depends on the type of material used, on the desired speed of operation, and on the particular system of pulsing used for selection. If the double pulsing system described by Rajchman is employed, the selection ratio need not be as high as otherwise, since the remanent point reached in this case is closer to the original state than if the negative swing were not present. The speed of the selection operation may be increased by choosing a value of $p>2$, and allowing the selecting field to exceed the coercive force considerably. These factors must all be weighed and a particular ratio selected. Then the actual set of wires to be used can be chosen by the criterion that the distances a and b which generate the lines be small.

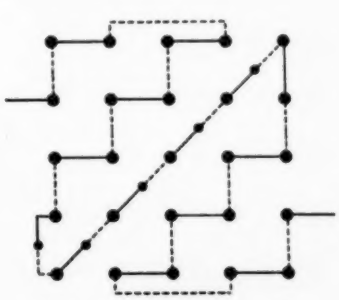


FIG. 4. Alternate method for threading the output wire.

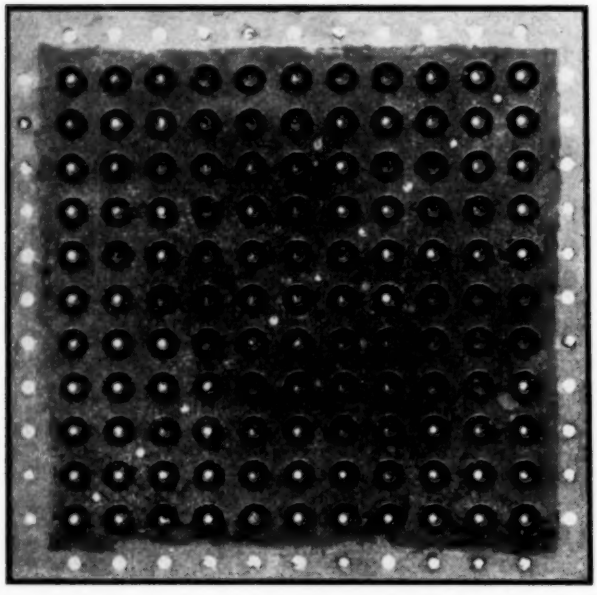


FIG. 5. One lamination of an 11 × 11 etched matrix.

It should be pointed out that using the present technique for selection considerably magnifies the problems of external switching. However, in some circumstances it is believed that the advantages in reliability and economy inherent in this multiple coincidence system offset the additional complexity of the control circuitry.

The problem of small signals accumulating on the output wire will now be considered. The first step in surmounting this difficulty is to thread the output wire so that it sometimes aids and sometimes opposes the polarities of the cores, and to rely on cancellation of the disturbance effects. If only the horizontal and vertical sets of wires are present, the output wire can be threaded in a fairly straightforward manner so that the number of cores aiding is practically equal to the number opposing. In the standard practice for discrete cores the output wire is run diagonally as shown in Fig. 3. All cores on each diagonal have the same polarity, either aiding or opposing; so that on any horizontal or vertical line the polarities alternately aid and oppose.

It is convenient to introduce the term *unbalance* as a measure of the total effect of the disturbed cores on the common output wire. The unbalance associated with a given core is defined as the difference between the number of disturbed cores which produce positive signals on the output wire and the number which produce negative signals when the given core is selected. It is easily seen that in the ordinary matrix system the range of unbalance is ± 2 if n is even, and ± 4 if n is odd. If the matrix is in the form of a perforated sheet, however, there is the additional constraint that the output wire must be threaded alternately in and out of the sheet, and since the input wires are also subject to this constraint, the pattern is more complicated. If the output wire is threaded in this manner along the path shown in Fig. 3 the unbalance ranges between ± 6, with respect to the sets of horizontal and vertical wires only. If, however, additional sets of wires are introduced to increase the selection ratio, the selection will involve more than two wires, and the additional wires will contribute to the unbalance. If the extra families of wires are those denoted by (1,2) and (2,1) then it can be shown[12] that the maximum range is ± 14. As mentioned earlier, these two sets always may be used if the least prime factor of n is equal to or greater than five.

Although the unbalance is large for the case just described, it does not depend on n, and so for large matrices the cancellation might still be expected to be relatively good. There is another system for introducing the output wire, however, which gives an unbalance of ± 1 per set of input wires when n is even and 0 or $+2$ per set when n is odd. This method of threading the wire is illustrated for odd values of n in Fig. 4, where the wire ordinarily is threaded in and out of alternate holes, but it is threaded through each core along the minor diagonal in the same direction. This is made possible by introducing small holes in the sheet represented by smaller dots in the figure. These holes do not represent cores but merely allow the wire to be threaded through successive cores in the same sense. This system gives the stated unbalance for any wire except one for which $\lambda - 1$ and n are relatively prime. Thus it is not satisfactory for the (1,1) set, but this set cannot be threaded in any case.

Even if there are as many cores aiding as opposing, the disturbing effects will not cancel exactly because of small differences in the cores themselves and because the disturbed response of a core differs, depending on

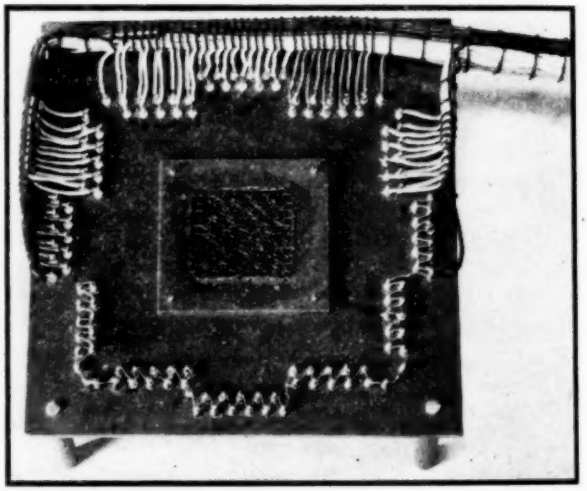


FIG. 6. Experimental 11 × 11 etched matrix.

[12] R. C. Minnick, "Output wire considerations for pierced multiple-coincidence magnetic storage systems," Harvard Computation Laboratory, *Bell Laboratories' Report Nos. 7 and 9, Theory of Switching* (May, October, 1954).

whether it stores a zero or a one. It should be noted that the magnitude of these effects is reduced considerably by increasing the selection ratio.

While it has been shown experimentally that the introduction of selection ratios greater than two makes it feasible to use pierced sheets of magnetic material as described above, it is also true that because of problems of flux linkage the holes must be spaced a considerable distance apart, which wastes space and magnetic material. Furthermore, the magnetic field about the wires through a given hole varies inversely as the radial distance from the hole. Thus if the magnetic material is continuous, it may be expected to appear much less square than if the same material were fashioned into discrete cores. It is still possible, however, to obtain systems which do not require the handling of individual cores in their construction. One such system which has been shown experimentally to be feasible consists of a bakelite plate with a square grid of holes drilled in it, to which is cemented a piece of magnetic material. By the use of standard photo-engraving techniques the material may be etched so that toroids of magnetic material are left around the holes. In Fig. 5 is shown an 11×11 array of cores produced from a two inch square sheet of 0.001 inch 4–79 molybdenum permalloy by this etching process. The distance between centers is 0.150 inches.

By stacking together several sheets, such as those in Fig. 5, a laminated structure is obtained. This not only lowers the switching time for a fixed cross-sectional area of the toroids, but it also tends to reduce variations in the magnetic properties of individual cores. A matrix made by assembling four such laminations is pictured in Fig. 6. This matrix was operated with selection ratios in the range $2 \leq p \leq 6$, and was shown[13] to operate reliably with $p \geq 4$.

[13] R. C. Minnick, "Etched magnetic storage system," Harvard Computation Laboratory *Bell Laboratories' Report No. 9, Theory of Switching* (October, 1954).

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Nucleation of Ammonium Iodide Crystals from Aqueous Solutions

J. B. NEWKIRK AND D. TURNBULL
General Electric Research Laboratory, Schenectady, New York

(Received September 17, 1954)

Ammonium iodide crystals nucleate at a sensible rate in small drops (10–50 micron diameter) of solution only if the supersaturation ratio S equals or exceeds a value of 1.21. This critical supersaturation ratio corresponds to an interfacial energy σ between ammonium iodide crystals and solution equal to or greater than 15.4 ergs/cm². However, for sensible nucleation of NH_4I on a freshly cleaved surface of one of the mica minerals, S is related to the disregistry δ between atom sites on the mica surface and the atom sites on a {111} plane of NH_4I. This dependence can be described by the equation $\ln S = 3.8\delta^2 + 0.0059$. This equation is in agreement with the theory of Turnbull and Vonnegut relating disregistry to the potency of nucleation catalysts. However, the coefficient of δ^2 is about 1 to 2 orders of magnitude smaller than the theoretical expectation.

Our results are consistent with Bradley's hypothesis that S, for sensible 2-dimensional nucleation of NH_4I on its own perfect {111} planes, is remarkably small, *viz.* $S \simeq 1.01$.

INTRODUCTION

TURNBULL and Vonnegut[1] have recently developed a theory which attempts to account for known facts concerning the catalysis of crystal nucleation by solid substances. The theory predicts that the order of potency of various catalysts will be inversely proportional to the atomic disregistry with certain low-index planes of the forming crystal.

Very little quantitative data are available by which this theory may be tested although much of the available evidence appears to fit qualitatively.[2,3,4] The purpose of the experiment described in this paper was to look for a quantitative relationship between the kinetics of nucleation and the atomic disregistry on the unstrained conjugate planes of the catalyst and the nucleus.

EXPERIMENTAL METHOD

It is well known that ammonium iodide, when crystallizing from aqueous solution in the presence of a freshly cleaved surface of the mineral mica, will form oriented pyramids of the solid salt upon the surface of the mica. This phenomenon is called "epitaxy" or "oriented overgrowth" and is a consequence of the similarity of atomic arrangement in the conjugate planes, in this case the cleavage plane of mica and a {111} plane of ammonium iodide. Although the arrange-

[1] D. Turnbull and B. Vonnegut, Ind. Eng. Chem. **44**, 1292 (1952).
[2] L. G. Schulz, Acta Cryst. **4**, 483–486 (1951).
[3] Yang, Birchenall, Pound, and Simnad, Acta Met. **2**, 462–469 (1954).
[4] Yang, Simnad, and Pound, Acta Met. **2**, 470–475 (1954).

TABLE I. Potency of different micas for nucleating ammonium iodide crystals in water solution.

Type of mica	Percent disregistry against (111) of NH_4I	Critical percent supersaturation at approx 30°C Average for several tests
Muscovite	1.16	0.66
Lepidolite	1.54	0.68
Zinnwaldite	2.66	0.85
Lepidomelane	3.40	1.07
Biotite	3.40	1.17
Phlogopite	3.76	1.05
Synthetic Phlogopite	3.76	1.20

ment of atoms on these planes is similar, the interatomic distances in ammonium iodide are slightly smaller than those of corresponding atom sites in mica, the actual amount depending upon the type of mica. Therefore, a crystallite of ammonium iodide which is coherent with a mica surface must be strained an amount not exceeding the disregistry of the unstrained conjugate planes.

By measuring the amount of undercooling which is necessary to form a nucleus of ammonium iodide on a particular type of mica, the energy required for nucleation may be related to the disregistry. Bradley[5] reported a critical undercooling of 1.45°C for the nucleation of NH_4I, on an unspecified type of mica, in a solution which was saturated at 31.5°C. He assumed that the strain imposed on the nucleus was zero and that the frequency of two-dimensional nucleation of ammonium iodide becomes sensible when the critical undercooling is equaled or exceeded.

Ammonium Iodide Structure

Ammonium iodide has the NaCl-type structure.[6] The {111} planes form alternate layers of ions of one type, either ammonium or iodide, arranged in hexagons. At room temperatures each atom in a {111} plane has three neighbors at a distance of $7.24/\sqrt{2}=5.12$A.

Mica Structure

Mica, having a space group $C^6_{2h}=C2/c$ and containing at least five different elements, has a much more complicated structure.[7] For the purpose of this experiment, the structure may be thought of as a succession of layers, unit groups of which are separated by parallel sheets of potassium atoms arranged in hexagonal array. This plane is also the cleavage plane of mica. Thus, a freshly cleaved surface of mica would be expected to consist of an hexagonal array of lattice sites, half of which are occupied by potassium atoms and half are vacant. The distance between these lattice sites varies with the type of mica. Since the atom distribution, on planes other than the cleavage plane, bears no simple relation to the atom distribution on any plane in an ammonium iodide crystal, one would not expect mica to catalyze an NH_4I nucleus on any except the mica cleavage plane.

Heterogeneous or Catalyzed Nucleation

If a nucleus of amonium iodide is to form on the surface of mica with $\{111\}_{NH_4I} /\!/ (001)_{mica}$ and $\langle 110\rangle_{NH_4I} /\!/ \langle 100\rangle_{mica}$, there will be a definite amount of misfit depending on the type of mica. For all known micas the a-parameter is larger than $(a_0/\sqrt{2})_{NH_4I}$ so that if the interface between ammonium iodide and mica is to contain no dislocations (i.e., the NH_4I forms coherently on mica) the lattice of the ammonium iodide must stretch or the mica must be compressed. The amount of this distortion can be expressed as the disregistry between the two undistorted lattices. If we assume that all distortion occurs in the ammonium iodide, we may define percent disregistry as:

$$\text{percent disregistry } (100\delta)=\frac{a(\text{mica})-\dfrac{a_a(NH_4I)}{\sqrt{2}}}{a(NH_4I)}\times 100.$$

Values of 100δ for ammonium iodide against the various micas are given in Table I.

A slight modification of the technique described by Bradley was used to measure the undercooling required to cause nucleation. Two clean glass bottles of about 25 ml capacity were arranged, as sketched in Fig. 1, with a communicating glass tube joining them through rubber stoppers. Bottle A, which was initially about three-quarters full of distilled water plus an excess of chemically pure ammonium iodide crystals, was fitted with a second glass tube which allowed the air pressure within the bottles to be controlled. In bottle B, a freshly cleaved sheet of mica was fastened to the

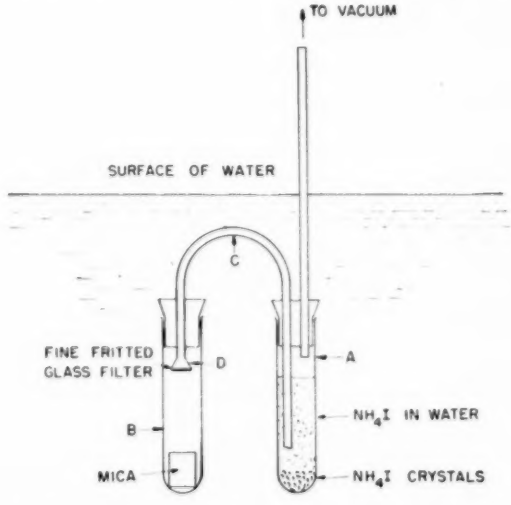


FIG. 1. Apparatus for determining undercooling required to nucleate ammonium iodide crystals on mica. (After Bradley.)

[5] R. S. Bradley, J. Chem. Soc. Part IV, 4530–4532 (1952).
[6] R. W. G. Wykoff, *Crystal Structures* (Interscience Publishers, Inc., New York, 1948), Vol. 1, Chap. III, Table page 21.
[7] W. L. Bragg, *The Atomic Structure of Minerals* (Cornell University Press, Ithaca, 1937).

bottom with an organic resin cement. The whole assembly was immersed in a water-bath at about 60°C and agitated so that the ammonium iodide crystals dissolved. The assembly was then transferred to a constant temperature bath, held at 30°±0.1°C, and was agitated for at least an hour after which time the solution was assumed to be saturated. This saturated solution was transferred to the bottle *B* by reducing the air pressure through tube *C* and then readmitting air. A fine fritted-glass filter (*D*) was included in the transfer tube to prevent any solid crystals from being carried over into bottle *B*. After sufficient solution had transferred to cover the mica, the bottles were removed from the water-bath and were separated. Bottle *B* was immediately closed with a clean rubber stopper and the solution in it was superheated by holding for 30 minutes at about 60°C. Finally the bottle was placed in the constant temperature bath held slightly above 30°C. Thereafter the temperature of the bath was lowered at a rate not exceeding 1°C per hour while the operator watched the mica surface carefully with a 10× magnifying glass. The temperature at which the first crystal appeared was taken to be the temperature at which nucleation occurred. This procedure is justified because of the low cooling rate and the rapid growth rate made possible by the available large volume of supersaturated solution.

Nucleation in Small Drops of Solution

It is well known that the nucleation of the first crystals in an undercooled liquid is usually catalyzed by foreign crystallites ("motes") of unknown origin and character. In principle the catalytic action of these accidental motes can be limited to a small fraction of the undercooled liquid if the liquid is broken up into noncommunicating droplets whose number greatly exceeds the number of motes. This "droplet technique" has been used effectively to elucidate the mechanism of nucleus formation in the crystallization of pure liquids. Recently White[8] has used the technique in investigating crystallization of supersaturated potassium nitrate solution.

In our investigation, chemically pure ammonium iodide crystals were dissolved in distilled water until a solution was obtained which was saturated at about 35°C. This solution was dispersed as small drops in warm silicone oil by beating with a high speed rotary mixer. Most of the droplets so produced were between 5 and 50 microns in diameter. A drop of this dispersion was placed on a glass slide, covered with a thin glass slip, and was examined microscopically. The temperature of the emulsion could be controlled by means of the apparatus shown in Fig. 2.

At low temperatures most of the droplets contained one or more crystals. These crystals dissolved as the temperature was raised and could be made to reappear by lowering the temperature again.

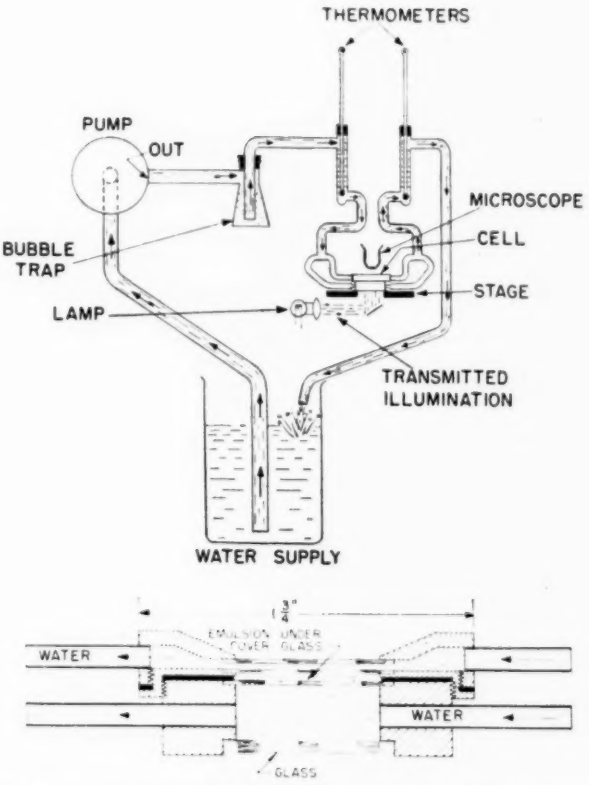


FIG. 2. Apparatus which allows the control of specimen temperature while the specimen is under microscopic examination.

During the emulsifying operation, visible "steam" was formed over the mixture, indicating that a considerable amount of water was evaporating from the droplets. This resulted in varied and unknown concentration differences among the droplets. Therefore the saturation temperature of each droplet had to be determined by heating slowly (0.2°C per minute) until the crystal within the drop disappeared.

The procedure generally followed was to heat the emulsion slowly, taking photographs every 2°C of a suitable field at a magnification of 25×. Heating was stopped at 55°C and the temperature of the emulsion was lowered to 35°C in about 5 minutes. Thereafter the cooling rate was 0.5°C per minute and a photograph was made every 2°C. The lowest temperature reached in the present tests was 4°C. After the film was developed it was examined on a microfilm reader where the disappearance and formation of a crystal within a droplet could be easily seen. This procedure made it possible to determine the saturation temperatures and the crystallization temperatures of each droplet to an accuracy of ±1°C. Therefore the undercooling was known to an accuracy of ±2°C. The undercooling required to form a crystal was recorded for about 30 droplets per test and four tests were made.

[8] M. L. White, Ph.D. dissertation, Department of Chemistry, Northwestern University, Evanston, Illinois (August, 1953).

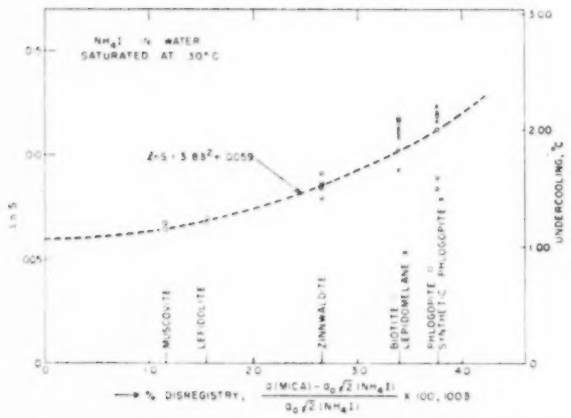


FIG. 3. Graph of percent disregistry between a {111} plane of ammonium iodide and the {001} plane of mica *vs* log of supersaturation ratio S required to cause nucleation at 30°C.

RESULTS

Heterogeneous Nucleation on Mica

The results for several tests of each mica are summarized in Table I and Fig. 3. It appears that the more closely the interatomic spacing on the mica surface approaches that on the {111} plane of ammonium iodide, the more potent is the mica in nucleating an ammonium iodide crystal. The data can be described by the following parabolic equation relating the disregistry (δ) with the logarithm of the supersaturation ratio (S) required to initiate nucleation:

$$\ln S = 3.8\delta^2 + 0.0059 \tag{1}$$

where

$$S = m/m_e$$

and

m = moles of NH_4I in solution per 1000 g of H_2O
m_e = moles of NH_4I per 1000 g H_2O in solution in equilibrium with crystalline NH_4I.

Nucleation in Small Drops

No consistent correlation of droplet size and critical undercooling was found. At undercoolings of 10, 20, and 30°C crystals had formed in about 5, 30, and 70 percent of the observed droplets respectively. Since the maximum undercooling found in any droplet was 34°C($S=1.21$), the formation of crystals at less than 34°C undercooling was presumed due to the presence of dirt particles which promoted nucleation. The experiments were not sufficiently extensive to establish that the critical undercooling required for a sensible rate of homogeneous nucleation is just 34°C. It can only be asserted that this critical undercooling equals or exceeds 34°C.

DISCUSSION

According to nucleation theory the change in free energy (ΔG) which accompanies the formation of a nucleus of a new phase is expressed by:

$$\Delta G = \text{volume of nucleus } (\Delta G_V) + \text{surface area } (\sigma),$$

where ΔG_V is the change in free energy per unit volume of new phase and σ is the interfacial energy between the nucleus and the parent phase. If the nucleus is strained, ΔG is increased by an amount proportional to the square of the strain, the elastic modulus (c), and the nucleus volume.

Assuming that the nucleus forms (coherently with the mica surface) as a right circular cylinder of radius r and height λ, we may write:

$$\Delta G = \pi r^2 \lambda \left(\Delta G_V + c\delta^2 + \frac{\beta}{2} \right) + 2\pi r \lambda \sigma, \tag{2}$$

where β is the change in free energy due to the difference in chemical interaction at the liquid-mica interface and the sum of the liquid-salt and the salt-mica interfaces, i.e.,

$$\beta = \sigma_{LM}' - (\sigma_{LS}' + \sigma_{SM}').$$

From this it follows that;

$$r^* = \frac{-\sigma}{\Delta G_V + c\delta^2 + \frac{\beta}{2}}, \tag{3}$$

and

$$\Delta G^* = \frac{-\pi\sigma^2\lambda}{\Delta G_V + c\delta^2 + \frac{\beta}{2}}, \tag{4}$$

where r^* is the radius of the critical size nucleus, σ is the interfacial energy between the nucleus and the solution,* δ is the disregistry or biaxial strain, and ΔG^* is the free energy change of the system when a critical nucleus forms. ΔG_V is given by:

$$\Delta G_V = -(kT/v) \ln(m\gamma/m_e\gamma_e)^2, \tag{5}$$

where γ is the mean ionic activity coefficient at the corresponding electrolyte concentration and v is the volume of an NH_4I "molecule" within the crystal. Unfortunately, we have not found data relating γ to the molality, m, covering the range of concentration and temperature used in our experiments. However, data for electrolytes comparable with NH_4I indicate that γ is not strongly dependent on m for the electrolyte concentrations used in our work. Therefore, we shall assume that $\gamma/\gamma_e \sim 1$ and

$$\Delta G_V = -(kT/v) \ln S^2. \tag{6}$$

The nucleation frequency I is related to ΔG^* by the general equation:

$$I = A \exp(-\Delta G^*/kT). \tag{7}$$

* We shall assume that σ, the edge energy for a two-dimensional nucleus of NH_4I on a mica surface is identical with the interfacial energy between a three-dimensional NH_4I nucleus and the solution.

A *minimum* value of σ can be estimated by applying the theory of homogeneous nucleation to our data on small drops. For spherical nuclei this theory requires:

$$\Delta G = 16\pi\sigma^3/3(\Delta G_V)^2. \tag{8}$$

Assuming that $I = 1$ cm^{-3} sec^{-1} at 34°C undercooling ($S = 1.21$), $A = 10^{30}$ cm^{-3} sec^{-1} (see reference 1), we calculate from Eqs. (6), (7), and (8) that $\sigma \geqslant 15.4$ ergs/cm^2.

For a sensible rate of nucleation, $I = 1$ cm^{-2} sec^{-1}, of NH_4I on the mica surface and with $A = 10^{22}$ cm^{-2} sec^{-1} for surface nucleation, it follows from Eqs. (4) and (7) that

$$\ln S = \frac{1}{2}\left[\frac{\pi\sigma^2\lambda v}{50k^2T^2} + \frac{\beta v}{\lambda kT} + \frac{cv\sigma^2}{kT}\right]. \tag{9}$$

Equation (9) has the same form as the empirical Eq. (1) that described the dependence of ln S on δ. Equating the corresponding constants in Eqs. (1) and (9) we have:

$$3.8 = \frac{cv}{2kT}, \tag{10}$$

and

$$0.0059 = \frac{1}{2}\left[\frac{\pi\sigma^2\lambda v}{50k^2T^2} + \frac{\beta v}{\lambda kT}\right]. \tag{11}$$

From Eq. (10) we calculate $c \approx 3.3 \times 10^9$ dynes/cm^2, a value about 100 times smaller than expectation based on the values of the bulk moduli of salts such as NaCl and KCl.

Setting $\lambda = 2.1 \times 10^{-8}$ cm (i.e., $\lambda = d_{111}/2$ for NH_4I) and $\sigma = 15.4$ ergs/cm^2, we obtain from Eq. (11) that $\beta = -0.05$ ergs/cm^2. Had we ignored β entirely, an edge energy of 13 ergs/cm^2 would have been calculated from Eq. (11). This edge energy agrees, within the experimental and theoretical uncertainty, with the minimum value of the interfacial energy between the three-dimensional nucleus and the solution, deduced from the small drop results.

If σ does not greatly exceed 15 ergs/cm^2, it follows that the frequency of two-dimensional nucleation of NH_4I on its own {111} planes must become sensible at the very low supersaturation ratio ~1.01 (corresponding to an undercooling of 1.5 to 2.0°) as Bradley has hypothesized. The free energy of the interface between vacuum and low index planes of NH_4I must greatly exceed 15.4 ergs/cm^2. However, the strong interaction between water molecules and the low index faces of NH_4I would be expected to greatly reduce the energy of these interfaces relative to their values in vacuo. The amount of this reduction is not known. If σ were 5 to 10 ergs/cm^2 greater than the minimum value which we have calculated, it would be impossible to initiate sensible homogeneous nucleation by cooling NH_4I solutions which were saturated at 40–100°C, without the prior formation of ice.

White's results are described with $\sigma \sim 45$ ergs/cm^2 between KNO_3 nuclei and water. Although this interfacial energy is three times greater than our minimum value for $NH_4I - H_2O$, the critical undercooling for sensible nucleation in small drops observed by White is about the same as we observed for NH_4I. The difference in σ values results from the fact that the temperature coefficient of solubility of KNO_3 is much greater than that of NH_4I so that at comparable undercooling the KNO_3 solution has a relatively much larger supersaturation ratio than the NH_4I solution. There is a possibility that σ may increase with increasing temperature coefficient of solubility.

Within the experimental scatter, the results relating the critical supersaturation ratio, S, for nucleation on mica to the disregistry between NH_4I and mica are in excellent qualitative agreement with the predictions of the theory of Turnbull and Vonnegut.[1] As predicted by this theory, S for the various micas is in the same order as the corresponding disregistries and ln$S$ varies as δ^2. However, this agreement between theory and experience cannot be taken very seriously at the present time because the experimental elastic coefficient is 20 to 100 times smaller than predicted. Perhaps more important that the quantitative slope of the ln$S$ *vs* δ relation is the new qualitative evidence that the potencies of various nucleation catalysts are inversely related to their disregistry with the nucleus. It is noteworthy that the cleavage planes of all of the micas in these experiments were within 4 percent of perfect registry with the {111} plane of NH_4I and all reduced the critical undercooling required for sensible nucleation of NH_4I from 34° to 2.5° or less. On the other hand, some qualitative observations that we have made indicate that cadmium iodide solutions can be undercooled about 50°C (corresponding to $S = 1.23$) without the formation of crystals on a mica cleavage plane. The disregistry between atom sites on the basal plane of CdI_2 crystals and on the cleavage face of mica is approximately 18 percent.

ACKNOWLEDGMENTS

Our thanks are hereby expressed to W. G. Martin who performed most of the experimental work described in this paper.

We wish to thank Mr. W. F. Foshag of the Smithsonian Institution for a sample of Zinnwaldite which he kindly supplied.

Self-Propagating Intermittent Discharge

W. S. BOYLE
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received July 31, 1954)

When a constant current pulse is applied to a pair of electrodes at small separations in air a series of electrical discharges takes place. These are characterized by a fixed distance between successive discharges. The intermittent nature of the discharge is determined by the local circuit and the spatial distribution of a series of discharges results from the blast wave which is propagated from each discharge. Measurements of the propagation distance as a function of the time interval between discharges and the energy released in each discharge are in qualitative agreement with those predicted from the theory of an intense explosion. The voltage transients across the electrodes show that during a series of discharges the breakdown voltages remain substantially constant and equal to the minimum breakdown voltage in air. This is attributed to the fact that the discharge can always find a low-density region which is quite near the front of the blast wave where the conditions for minimum breakdown voltage can be satisfied.

THE electrical discharge patterns shown in Fig. 1 and Fig. 2 are the result of applying a single short duration constant current pulse across the gap between separated electrodes. In Fig. 1, one electrode is a smooth flat surface and the other is a long straight wire placed parallel to this surface; in Fig. 2 the wire has been replaced by a conical electrode placed close to a flat surface. The white spots in the two figures appear on the cathode and each represents a single small arc. During the course of the current pulse there has in each case been a number of discrete arcs occurring one after another in different places. The intermittent nature of the discharge results from the local circuit associated with the electrodes behaving as a relaxation oscillator. The ordered spatial arrays of discharges are consequences of shock and blast waves which are propagated out from the discharges. Each discharge takes place at that point in the blast wave from the previous discharge which best satisfies the electrical breakdown conditions.

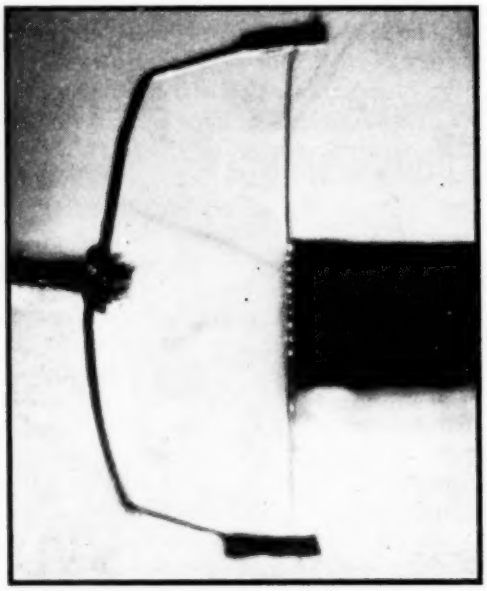


FIG. 1. An ordered array of discharges between a wire cathode parallel to a flat anode.

The density distribution in a gas following an infinitely concentrated release of energy in an infinitely short time has been calculated for the case of spherical symmetry.[1] When the shock wave generated by the explosion is sufficiently intense a density distribution as shown in Fig. 3 is obtained. It is to be noted that the distribution is dimensionless, and so long as the pressure jump across the shock front is sufficiently large (>10), this distribution remains the same. A rough estimate of the energy released during one of the discharges photographed in Figs. 1 and 2 shows that for a large fraction of the time interval between successive discharges the shock front is sufficiently intense to apply the above theory, and even at the end of the interval this density distribution is approximately correct.

An approximate formula is given by Taylor for the position of the shock front in the expression

$$t=0.926R^{5/2}\rho^{1/2}E^{-1/2}, \qquad (1)$$

where t = time since the explosion, R = radial coordinate

FIG. 2. Luminous spots on a conical cathode during a single current pulse.

[1] G. I. Taylor, Proc. Roy. Soc. (London) **A201**, 158 (1950).

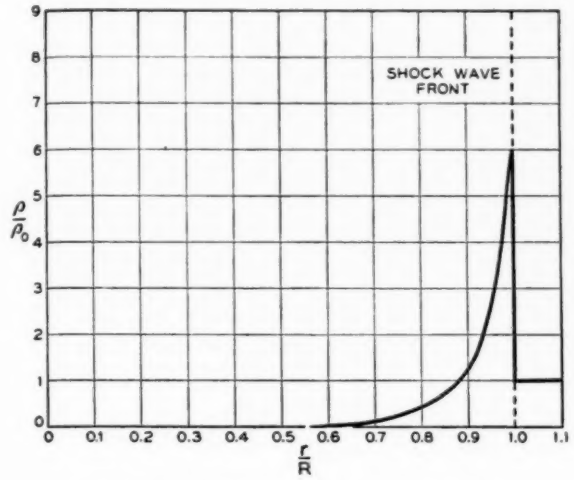
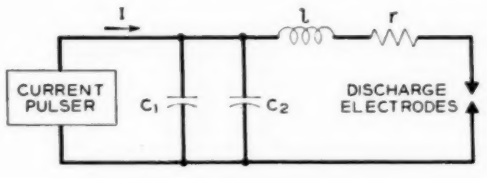


FIG. 3. Radial distribution of the density in an intense explosion of spherical symmetry.

of the shock front, ρ=density of gas surrounding the explosion, E=total energy released.

The electrical breakdown voltage in a gas (V_B) is usually expressed by a curve where V_B is plotted against pd, the pressure and separation of the electrodes. Since in this case the temperature is not constant throughout the gas, it is more appropriate to consider V_B *vs* ρd. It is evident from the density distribution (Fig. 3) that a value of ρd corresponding to the minimum sparking potential will always be satisfied at some point behind the shock front and that for small electrode separations this point will be quite near the front. The breakdown will be affected also by the ions carried along by the large fraction of the gas which follows closely behind the shock front. These ions will be effective in reducing the statistical time lag and will also tend to reduce the effective separation between the electrodes by their space charge and thus act in the same sense as the decreasing density distribution.

The experimental results presented below were obtained using a hard-tube current pulser which delivered square-wave current pulses of 1→7 amperes for 1→4 microseconds. A schematic diagram of the circuit associated with the discharge electrodes is shown in Fig. 4; l is the small inductance of the loop



C_1 STRAY CAPACITY OF LEADS AND PULSER TUBES
C_2 ADDED LUMPED CAPACITY
l INDUCTANCE OF DISCHARGE LOOP
r RESISTANCE OF DISCHARGE LOOP

FIG. 4. The circuit associated with the electrical discharge shown in terms of lumped parameters.

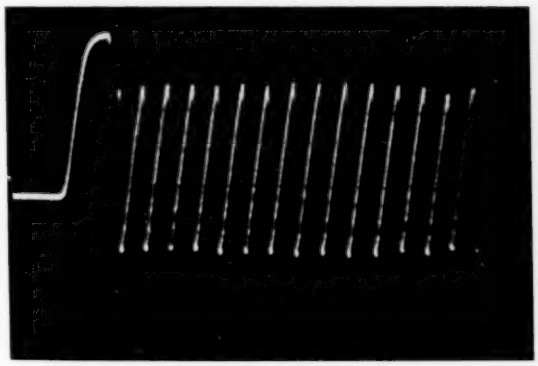


FIG. 5. Voltage transients across the parallel capacity during an intermittent discharge. Capacity 200 $\mu\mu$f. Pulse current 3.0 amperes. (The amplifier is overloaded on the first breakdown which occurs at 900 volts.)

leading to the electrodes, and r is the resistance of this loop. The capacitance of the pulse tube and of the leads are included with a lumped additional capacity in the term $C_1+C_2=C$.

The intermittent nature of the discharge is readily understood in terms of this local Crl circuit. If the arc is extinguished when the current between the electrodes just reaches zero, it can be easily shown that the necessary condition that the current be intermittent is that

$$(V_B^2C/I^2l+1)^{\frac{1}{2}}\exp[-(r\pi/2)(C/l)^{\frac{1}{2}}]>1 \qquad (2)^*$$

If the pulse current exceeds the critical value given by (2), a steady metal arc develops.

The voltage transients across the lumped parallel

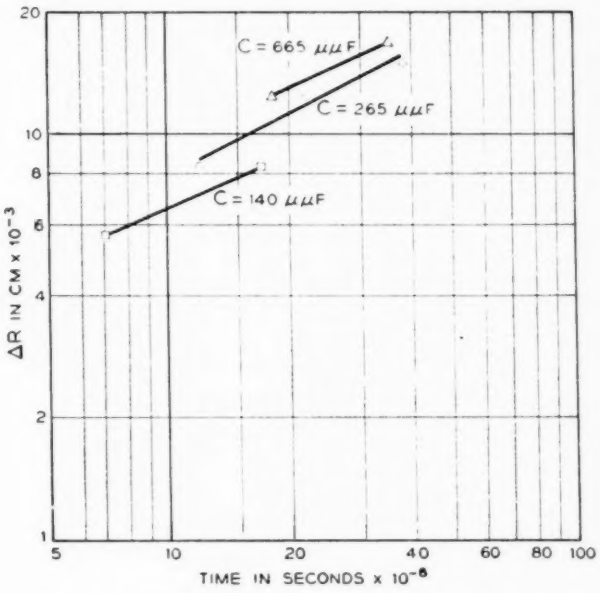


FIG. 6. Separation of successive discharges as a function of the time interval.

* This equation is derived by assuming that the initiation of the arc can be represented by an instantaneous drop of the electrode voltage from V_B to zero.

TABLE I.

C $\mu\mu f$	I Amperes	R 10^{-3} cm	t μ seconds	"Velocity" m/sec
140	0.6	8.3	0.17	490
140	1.6	5.7	0.07	820
265	0.5	16.0	0.38	420
265	0.9	12.5	0.22	570
265	1.9	8.4	0.12	700
665	1.6	17.0	0.35	480
665	3.0	11.7	0.18	650

capacity C_2 are shown in Fig. 5. The first breakdown voltage will depend on the electrode separation: for 50 000 A this is about 900 V for gold electrodes in air but varies by some 300 volts because of the statistical time lag of breakdown. All subsequent breakdowns occur at 340 volts. The remarkably constant value of breakdown voltage after the first high value is evidence that the pd value for the minimum sparking potential has been reached; the experimental value of 340 volts is in fairly good agreement with the minimum breakdown voltage in air.

A series of measurements of the separation between such spots as those shown in Fig. 1 was obtained for various values of the pulser current and parallel capacity C. The results are presented in Table I along with the time interval as obtained from the oscilloscopic trace of the voltage transients across the condenser.

A plot of $\log R$ *vs* $\log t$ values from Table I is given in Fig. 6. According to (1) the slope of these curves should be 2.5; the values obtained here are of the order of 1.8. Recently a solution for the case of a cylindrical shock wave has been given,[2] and in this case the slope should be 2. This is perhaps more appropriate for our problem and the agreement with the experimental results is certainly better. The accuracy of the experimental results do not, however, warrant distinguishing between the two geometries. For a given value of t, we have from (1) $R \propto E^{1/5}$. Since E is proportional to C we find that this formula also is in good qualitative agreement with the curves in Fig. 6 for the three experimental values of the capacity.

It is evident that the shock-wave hypothesis of the self-propagating discharge is well substantiated by the experimental results on the velocity of propagation. The fact that all breakdown voltages occur at constant value which is the minimum breakdown voltage in air lends further support to this mechanism.

[2] Shao-Chi Lin, J. Appl. Phys. **25**, 54 (1954).

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Signal-to-Noise Improvement and the Statistics of Track Populations*

NELSON WAX†
University of Illinois, Urbana, Illinois
(Received September 29, 1954)

A treatment of the population statistics of signal and of noise patterns, or tracks, is given in this paper. Some idealized and abstract models of tracking procedures are studied; the models are intended to represent the processes encountered in radar tracking and in scanning nuclear emulsions for tracks. The equilibrium track population, the age distribution of the population at equilibrium, and a signal-to-noise improvement factor are obtained for each of the models considered. It is shown that the equilibrium population can be kept to reasonably small values, and the signal-to-noise ratio improved, if effective methods are used to eliminate deteriorating tracks.

1. INTRODUCTION

THE detection of signals, and in particular the detection of radar signals in the presence of noise has been the subject of numerous investigations. Early studies of radar signal detectability often incorporated the role of the human observer into the theory,[1] but it was soon felt that a fundamental approach to understanding the statistical questions involved was necessary. Some later work[2] has therefore omitted considerations of human observation and concentrated on detectability criteria which are precisely defined.

One important function of the human observer has thereby been neglected: namely, the recognition of "patterns" or tracks which may be formed from a set of observations. For example, the major duty of a radar operator is to recognize tracks in a set of "blips" on the PPI screen. The "blips" may be the record of radar returns from a moving aircraft, they may be a representation of a random collection of unrelated events, or, very often, the "blips" are a combination of both occurrences.

* The research reported in this paper was made possible by support extended to the University of Illinois, Control Systems Laboratory, jointly by the Department of the Army (Signal Corps and Ordnance Corps), Department of the Navy (Office of Naval Research), and the Department of the Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps contract.

† On leave as John Simon Guggenheim Memorial Foundation Fellow at the Chalmers Institute of Technology, Gothenburg, Sweden.

[1] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals*, Massachusetts Institute of Technology, Radiation Laboratory Series (McGraw-Hill Book Company, Inc., New York, 1950), Vol. 24.

[2] D. Middleton, J. Appl. Phys. **24**, 371 (1953); **24**, 379 (1953).

The scanner of the plates of nuclear emulsions again must recognize patterns in a set of grains. The developed grains in the emulsion now may represent the track of a nuclear particle, they may be samples of "noise," or of tracks and "noise." The detector in this case is the nuclear emulsion and the developing process; the observations are the observed positions of the developed grains. Noise arises both as a background of single, or small collections of isolated grains, and also occurs in the random size of the grains making up the tracks of actual nuclear particles.

Similar needs for pattern recognition exist in a variety of other physical situations but even when the patterns themselves are not of primary interest, it still may be useful to search for regularity in a set of observations. The pattern may be of help in correcting observations which were labelled erroneously during the detection process. Discussions of pattern formation and recognition therefore seem to be desirable.

General, and somewhat abstract models of pattern formation, pattern recognition, and pattern rejection will be treated in this paper. (The models are intended to portray, in a highly idealized and simplified fashion, the processes encountered in radar tracking, in scanning nuclear emulsions for tracks, and in other applications.) The particular patterns to be studied are either the tracks which can be formed from the records of the motion of real objects, whose positions are recorded at discrete time intervals,‡ or similar patterns which may arise from noise. In the first case the patterns will be termed real tracks, actual tracks, or signal tracks, interchangeably; the phrase "noise tracks" will be reserved for the second possibility.

Certain elements of track formation (birth), track maintenance (life), and track rejection (death) appear to be common to a number of applications, and seem to be essential enough to be included in any model. These elements are a memory, where the observations are kept, a set of rules for accepting or rejecting a set of observations as forming a track and a computing facility which applies the rules to the observations. Thus the photographic plate serves as a memory unit for the scanner, who applies his own (usually unformulated) rules to the data in making a decision that a real track exists.

A distinction has been made between the detector, a device for recognizing the presence or absence of a possible signal (e.g., in the nuclear emulsion problem, recognizing a grain) and another data processor, whose function is to recognize the presence or absence of a track in the set of possible signals (grains). The detector therefore is to act as a filter of the original data, selecting only the possible signals; the data processor is to filter the possible signals for tracks. The distinction between detectors and data processors is frequently an arbitrary one, since some devices combine both functions, and contain memory elements, a set of built-in rules, and computing organs. It is the size of the memory which often governs the choice between using a single unit, which performs both detection and pattern recognition, and devices where the operations are separated. The separate detector and data processor frequently require a smaller total memory capacity than is needed by one combined processor.

Further discussion of memory size and of the related statistical questions will be found in the next section. The assumptions, on which the models are based, are listed there, too. Birth processes are the subject of Sec. 3. One model of birth and death is studied in Sec. 4, another in Sec. 5. Section 6 contains a general treatment of birth and death processes. Signal-to-noise improvement is investigated in Sec. 7. The results are discussed in Sec. 8.

2. GENERAL CONSIDERATIONS

The system which is to be treated in this paper is one which contains one or more sense organs, their detectors, and some sort of computer; the primary function of the system is to form tracks which are representations of the trajectories of moving objects. The trajectories are taken to be segments of smooth simple curves.

Specific assumptions concerning the operation of the sense organs, the detectors and the computer are listed in this section. The listing will proceed systematically from the sense organs to the computer, whenever it appears feasible. In addition, the statistical questions which arise in forming or deleting a track are discussed briefly here; further details are given in the Appendix.

A. Assumptions

(1) *Phase Space Quantization*

The resolution of the sense organs sets limits to the position in space to which an object can be assigned. Further, the sense organs may be devices for determining velocity; hence limits to velocity assignments may exist. The "region of uncertainty" in phase space, set by resolution of instruments, will be termed an "observation bin."

It is assumed that the detector quantizes phase space; the quantized regions may be of the order of an observation bin size but whatever these bin sizes may be, the computer will use some characteristic region of phase space as a unit. This region, termed a "bin," is assumed to be of constant size.

(2) *Amplitude Quantization*

The sensitivity of the sense organs and of the detectors sets limits to the number of distinguishable classes into which the observed intensity or amplitude of a return can be put, for a given bin. It is assumed that there are but three such classes, the first corresponding to small or negligible amplitudes, for which no report is

‡ Time has been chosen as the independent variable; the choice of a space dimension might be more appropriate for the nuclear emulsion problem.

issued to the computer, the second corresponding to "medium" amplitudes, and the last to large amplitudes. In the latter two cases reports are furnished to the computer giving the amplitude classes to which the reports belong.

(3) *Report Rates*

The functions of the sense organs and of the detectors may be summarized by stating that the presence or absence of a possible object in any bin is noted and reported to the computer, together with the amplitude class of the possible signal observed by the sense organ.

It is assumed that every bin in phase space is examined periodically for possible signals. The period, or frame time, will be chosen as the unit of time. It is convenient, but not essential, to suppose that a constant number of reports arrive at the computer during each frame time. Further, it is assumed that the number of bins for which signal reports occur in each period, N_2, and the number of bins for which noise reports are issued per period, N_1, are constants, and that $(N_1+N_2)\ll N_T$, where N_T is the total number of bins in phase space.

(4) *Track Independence*

The computer uses a report either by starting a new track, continuing an already existing track, or by ignoring the information. Cases of ambiguity can, and in practice do, arise, namely where a report may be assigned to any one of a number of tracks which cross, or come close to each other. These ambiguities are ignored here. In effect, it is assumed that every track is independent of every other track, or, equivalently, that no track interacts with any other track.

(5) *Constant Association Region*

Suppose that a report initiates a (tentative) track at a given time. If velocity information, as supplied by the sense organs and the detectors, is part of the report, then the computer can perform a computation (according to some appropriately chosen formula) which will yield a prediction of the bin, or bins, where the next report could be expected to occur. If a report does arrive at a suitable time in any one of the predicted bins, then an association has occured and the track is continued. The new report, containing another observation of position and velocity, allows one to correct or smooth the velocity determinations and thus perform a more refined prediction computation. If reports continue to arrive at appropriate times, in suitable bins, then the track can be determined with increasing accuracy and the number of predicted bins decreases (or reaches some lower limit).

A similar decrease in the size of the predicted region, termed an association region, occurs even if velocity information is not part of a report. One has then to wait for two observations of position before a velocity can first be computed, but thereafter the prediction computations proceed much as before (although the detailed operations may differ) and the association region decreases in size.

If, however, one or a few reports are missing in the sequence, then it is reasonable to increase the size of the region of association, and accept, as a continuation of a track, any report which occurs at the proper time in the new region of association.

The size of the association region usually increases and decreases, therefore, depending on the history of the reports, the computational procedures, and the rules for track continuation. It will be assumed, however, that the association region is of constant size.

(6) *Constant Probabilities*

The *a priori* probabilities of obtaining reports for a given bin in either the medium or large amplitude classes differ for noise reports and for signal and noise reports. Furthermore, the probabilities are in general functions of the bin's location in phase space. It will be assumed, however, that each of the probabilities is constant over phase space.

(7) *Size of Memory*

After performing a prediction computation the computer stores the results, together with some (or all) of the previous track information in a memory, for future reference. If ν be the total number of tracks which exist at any given time, then the necessary size of the memory will be taken to be directly proportional to the maximum value of ν. The total track population should be kept small, therefore, if the memory is to be small.

B. Statistical Preliminaries

The two requirements, that tracks be identified correctly as being real or noise tracks, and that the memory be kept small, conflict somewhat, since greater reliability in making a decision is often attainable at the expense of increased memory capacity. A statistical test is desired, therefore, which will allow one to decide that a track is signal or noise early in the track's history, which will use additional reports in a re-examination of the original decision and which will permit the deletion of a track quickly at the end of the track's life-span. These functions are to be performed with small error, and, if possible, the test should be one readily amenable to computation.

The methods commonly used in statistics have not been designed with the above ends in view. A decision is reached concerning the presence or absence of a signal either after a fixed number of observations[3] (Neymann-Pearson test) or after a run of varying length[4] (Sequential test). In either case, however, additional observations are not used, once a decision is made.

[3] H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, New Jersey, 1946).

[4] A. Wald, *Sequential Analysis* (John Wiley and Sons, Inc., New York, 1947).

The tests used here, which are applications of known procedures, have been chosen largely for their computational simplicity. The specific details of the tests are given in the following sections, and in the Appendix.

3. BIRTH PROCESS

The initiation of a track is a somewhat arbitrary procedure, since it is a question of definition as to when a track "begins." Two different birth processes will be considered in this paper. In the present section the simplest possible criterion will be used for birth, namely that a sufficiently strong unassociated report constitutes the beginning of a new track; the track appears immediately as a member of the population. A more complicated view is adopted in later sections. There a track is subject to a period of gestation before emerging as a genuine member of the track population.

The assumption that tracks are independent allows one to treat the noise and signal tracks separately. Just one class of tracks will be considered, therefore, until signal-to-noise improvement is investigated.

(a) Single Bias Level

Rules for track initiation and maintenance: a track is started in the computer by any new, i.e., unassociated, report. A track once initiated, is kept by the computer thereafter; new track positions are obtained either by associations which may occur, or by continued extrapolation.

Let a be the constant number of bins in which a report can appear that will serve as an association region, $\nu(r)$ be the average§ number of tracks carried by the computer at time r, N be the constant number of reports per unit time, and p' the constant probability of a report appearing in any bin.

The birth equation may then be written as

$$\nu(r)=\nu(r-1)+[N-ap'\nu(r-1)], \tag{3.1}$$

which, in words, states that the average number of tracks at time r equals the average number at time $(r-1)$ plus the average number born during the last time interval. Those born are those unassociated, namely, $N-ap'\nu(r-1)$.

It is convenient to introduce the reduced variable, $n(r)$, the average number of tracks at time r per report,

$$n(r)=\nu(r)/N, \tag{3.2}$$

and the parameter

$$p=ap', \tag{3.3}$$

where p is the probability of an association.

Equation (3.1) may be written as

$$n(r)=1+(1-p)n(r-1). \tag{3.4}$$

The general solution of Eq. (3.4) with $n(0)=1$, $[\nu(0)=N]$, is given by

$$n(r)=\frac{1}{p}[1-(1-p)^{r+1}]. \tag{3.5}$$

For

$$r\to\infty,\quad n(r)\to 1/p,\quad \text{or}\quad \nu(r)\to\frac{N}{ap'}=\frac{N}{aN/N_T}=\frac{N_T}{a}.$$

Thus the limiting case implies that every available region of phase space contains a track; under these circumstances the analysis is no longer valid.

(b) Two Bias Levels

Rules for birth: a track is initiated by unassociated reports which are of large amplitude; tracks are continued by extrapolation or association as in case (a).

Let Q be the conditional probability that a report is of large amplitude, given that the amplitude of the report exceeds or equals the bias for medium amplitudes.

The birth equation may now be written as

$$n(r)=n(r-1)+Q[1-pn(r-1)], \tag{3.6}$$

with the general solution, setting $n(0)=Q$, given by

$$n(r)=\frac{1}{p}[1-(1-pQ)^{r+1}]. \tag{3.7}$$

As before $n(r)\to 1/p$ when $r\to\infty$. Thus the limits are the same in cases (a) and (b), merely being approached more slowly when $Q<1$.

Equation (3.7) yields an upper bound for the total track population found in the next and later sections.

4. BIRTH AND DEATH PROCESS—A

The criterion used above for birth is a simple and perhaps even a natural one; the death processes considered here are analytically more complicated than the birth process. The additional complication arises because it is reasonable to introduce the age distribution of the population, in computing the death rate.

A particular scheme for track initiation, maintenance, and elimination is described in the present section. Another scheme is investigated in Sec. 5.

The rules are: Birth—same as in Sec. 3(b). Maintenance—a track is continued, by extrapolation or association, until it is eliminated. Death—a track is eliminated only if it has suffered m successive nonassociations. The integer m is, for the moment, unspecified. Let $n(r;i)$ be the number of tracks at time r which have had i successive nonassociations. The notation used to indicate the age distribution is $n(r;i)$. A track dies if $i=m$. Thus

$$n(r)=\sum_{i=0}^{m-1} n(r;i). \tag{4.1}$$

Now the number of tracks with i nonassociations at

§ "Average" and "expected" are to be interpreted as "mathematical expectation" throughout this paper. The terms "average" and "expected" will be omitted frequently.

time r is equal to that fraction of the number of tracks at time $r-1$ which had $i-1$ nonassociations and were again not associated. Or

$$n(r;i)=(1-p)n(r-1;i-1). \tag{4.2}$$

Hence

$$n(r;m)=(1-p)^m n(r-m;0). \tag{4.3}$$

The birth and death equation can now be written as

$$n(r)=n(r-1)+Q[1-pn(r-1)] -(1-p)^m n(r-m;0), \tag{4.4}$$

or, using Eq. (4.1), as

$$\sum_{i=0}^{m-1} n(r;i)=Q+(1-pQ)\sum_{i=0}^{m-1} n(r-1;1) -(1-p)^m n(r-m;0). \tag{4.5}$$

Let $U(r)=n(r;0)$. Then, using Eq. (4.2), one may write Eq. (4.5) in the compact form

$$\sum_{i=0}^{m} b_i U(r-i)=Q, \tag{4.6}$$

where $b_0=1$, $b_i=-p(1-p)^{i-1}(1-Q)$, $1\leq i\leq m$. The general solution of Eq. (4.6) is given by

$$U(r)=\frac{Q}{Q+(1-Q)(1-p)^m}+\sum_{s=1}^{m}\omega_s x_s{}^r, \quad r\geq m \tag{4.7}$$

where the ω_s are given by the initial conditions, and $x_1, \cdots, x_m$ are the m roots of the homogeneous equation

$$\sum_{i=0}^{m} b_i x^{m-i}=0. \tag{4.8}$$

It follows that

$$n(r)=\frac{Q[1-(1-p)^m]}{p[Q+(1-Q)(1-p)^m]}+\sum_{s=1}^{m}\gamma_s x_s{}^r, \quad r\geq m \tag{4.9}$$

where $\gamma_s=[1-(1-p)^m](\omega_s/p)$.

The m values of γ_s can be determined from the set of m linear equations obtained by equating $n(r)$, given by Eq. (4.9), to $n(r)$, given by the birth process, Eq. (3.5), for $r=0, 1, \cdots, m-1$.

The quantity of greatest interest is the equilibrium population, $\lim_{r\to\infty} n(r)$, if it exists. The existence is established by noting that if x_1 be the largest real root of Eq. (4.8), then

$$\sum_{s=1}^{m}\gamma_s x_s{}^r\sim\gamma_1 x_1{}^r\to 0, \tag{4.10}$$

since $0<x_1<1$, if both p and Q are between 0 and 1.[5]

[5] L. E. Dickson, *Elementary Theory of Equations* (John Wiley and Sons, Inc., New York, 1914), Ex. 22, p. 168.

Hence

$$\lim_{r\to\infty} n(r)=\frac{Q[1-(1-p)^m]}{p[Q+(1-Q)(1-p)^m]}. \tag{4.11}$$

If $mp\ll 1$ then one has

$$\lim_{r\to\infty} n(r)\doteq mQ. \tag{4.12}$$

If one were interested just in keeping the total population small, then m should be small. Small values of m will result in a high death rate for real as well as for spurious tracks, however. A rational choice of m should depend, therefore on the statistics of both real and spurious tracks. One method of choosing m is outlined in the Appendix.

5. BIRTH AND DEATH PROCESS—B

The population considered in the last section contained only vigorous members: a track was stored in the computer's memory as being *bona fide*, or else it did not exist. Further discrimination between tracks is possible, and at times desirable, and such discrimination is considered in the present section.

In order to distinguish between tracks it is convenient to assign a number, f, to a track, on first appearance. The number f, called the firmness,‖ is to serve as a running count, or index, of one's degree of confidence that the track is an actual track; f will change during the track's history. The rules are:

(1) A track is initiated by a large amplitude report, as in Sec. 3(b). On initiation a track is assigned the integer f_I.

(2) On each succeeding association the index, f, is increased by the integer l (except near f_M—see rule (5) below); each nonassociation decreases f by 1.

(3) A track is eliminated when $f=0$.

(4) A track is accepted as an actual track when $f\geq f_A$.

(5) f has a maximum value, f_M. If $f_M-l\leq f\leq f_M$, then an association increases f to f_M (or leaves f unchanged, at $f=f_M$).

Note that a track is considered to be tentative when $0<f<f_A$. The maximum, f_M, has been introduced in order to limit the memory requirements.

The integers f_I, f_A, f_M, and l are determined, from the signal and noise track statistics, in the Appendix.

Let $n(r;f)$ denote the reduced number of tracks, at time r, with firmness f. Then, using the rules given above, one may describe the birth and death process by the sequence of difference equations:

Case (a) $f_I>l$

$$n(r;f_I)=pn(r-1;f_I-l)+(1-p)n(r-1;f_I+1) +Q[1-pn(r-1)], \tag{5.1}$$

$$n(r;f)=pn(r-1;f-l)+(1-p)n(r-1;f+1), \quad l+1\leq f\leq f_M-1 \tag{5.2}$$

‖ A term suggested by Dr. A. T. Nordsieck.

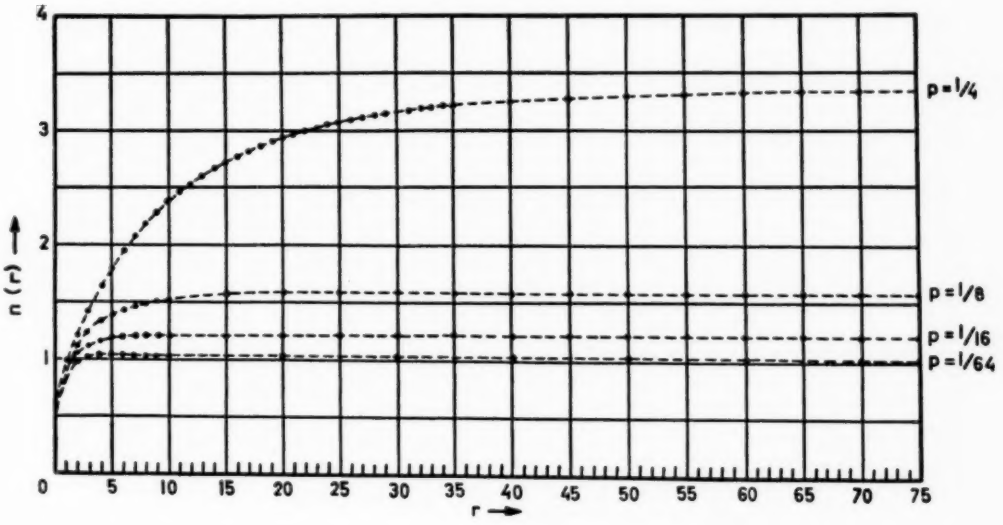


FIG. 1. The total track population in reduced units $n(r)$ *versus* r, the number of frame times for various values of p, when $Q=\frac{1}{2}$.

$$n(r; f_M)=p\sum_{f=f_M-l}^{f_M} n(r-1; f), \tag{5.3}$$

$$n(r; f)=(1-p)n(r-1; f+1), \quad 0\leq f\leq l \tag{5.4}$$

and

$$n(r)=\sum_{f=1}^{f_M} n(r; f). \tag{5.5}$$

Case (b) $f_I=l$

For this case, the aforementioned system of equations (5.1–5.5) needs to be modified slightly. Equation (5.1) should be replaced by

$$n(r; f_I)=(1-p)n(r-1; f_I+1)+Q[1-pn(r-1)], \tag{5.6}$$

and Eq. (5.4) is to hold now for all $0\leq f\leq l-1$. The remaining equations are still applicable.

Case (c) $f_I<l$

Equation (5.4) is to hold now when $0\leq f\leq f_I-1$. All the other equations of Case (b) are still applicable.

Each of the above systems may be reduced to a single equation by summing on f. One obtains

$$n(r)=Q[1-pn(r-1)]+n(r-1)-(1-p)n(r-1; 1), \tag{5.7}$$

where the first term in the right hand member of Eq. (5.7) represents the tracks born, the second term those that continue to live, and the last those that die.

Equation (5.7), or the appropriate system of equations, together with the initial conditions, is a complete description of the birth and death process. The initial conditions are taken to be

$$\left.\begin{aligned} n(0; f_I)&=Q,\\ n(0; f)&=0, \quad f\neq f_I\end{aligned}\right\}. \tag{5.8}$$

A treatment of systems with arbitrary parameters is given in the next section. It may be instructive, however, to devote the remainder of the present section to the examination of a particular process, first suggested by Dr. A. T. Nordsieck. Let $f_I=l=2$, $f_A=5$, $f_M=7$; then one has

$$n(r; 0)=(1-p)n(r-1; 1), \tag{5.9}$$

$$n(r; 1)=(1-p)n(r-1; 2), \tag{5.10}$$

$$n(r; 2)=(1-p)n(r-1; 3)+Q[1-pn(r-1)], \tag{5.11}$$

$$n(r; f)=pn(r-1; f+2)+(1-p)n(r-1; f+1), \quad 3\leq f\leq 6 \tag{5.12}$$

$$n(r; 7)=p\sum_{f=5}^{7} n(r; f), \tag{5.13}$$

and

$$n(r)=\sum_{f=1}^{7} n(r; f). \tag{5.14}$$

Solutions of the system (5.9–5.14), with the initial conditions (5.8), have been obtained for $Q=\frac{1}{2}$ and various values of p, using the University of Illinois Digital Computer. The solutions are shown in Fig. 1.

The quantities of greatest interest are the equilibrium values of $n(r; f)$, if they exist; $n(r; f)$ will then be independent of r for sufficiently large r.

Let

$$\left.\begin{aligned} \lim_{r\to\infty} n(r; f)&=w_f,\\ \sum_{f=1}^{7} w_f&=W.\end{aligned}\right\} \tag{5.15}$$

One has, then, the system of algebraic equations

$$w_1=(1-p)w_2, \tag{5.16}$$

$$w_2=(1-p)w_3+Q[1-pW], \tag{5.17}$$

$$w_f=pw_{f-2}+(1-p)w_{f+1}, \quad 3\leq f\leq 6 \tag{5.18}$$

$$w_7=p\sum_{f=5}^{7} w_f, \tag{5.19}$$

as the limiting form of the system of difference equations (5.10–5.14), when $r\to\infty$, if the w_i exist.

The solutions of Eqs. (5.16–5.19) are given by

$$w_i=\frac{\Delta_i}{D}, \quad i=1, \cdots, 7, \tag{5.20}$$

where

$$\Delta_1=\frac{(1-p)^6}{p}, \tag{5.21}$$

$$\Delta_2=\frac{(1-p)^5}{p}, \tag{5.22}$$

$$\Delta_3=(2-p)(1-p)^4, \tag{5.23}$$

$$\Delta_4=(1-p)^3+p(1-p)^4, \tag{5.24}$$

$$\Delta_5=p(1-p)^2(3-2p), \tag{5.25}$$

$$\Delta_6=p(1-p)^2+p^2(1-p)(2-p)^2, \tag{5.26}$$

$$\Delta_7=p^3(2-p)^2+2p^2(1-p)(2-p) \tag{5.27}$$

and

$$D=(1-p)^5\left[\frac{1+pQ}{pQ}+(1-p)\right]+p^2(2-p)$$
$$\times[1-p(1-p)^2]+p(1-p)[1-p(1-p)^2]$$
$$-p(2-p)(1-p)\left[\frac{1-p-pQ}{pQ}-p\right]. \tag{5.28}$$

Two values of W have been computed, using the above equations ($Q=\frac{1}{2}$ and $p=\frac{1}{8}$, $p=\frac{1}{16}$), and found to agree with the corresponding asymptotic values, shown in Fig. 1, to within 5 percent.

The w_i, given by Eqs. (5.20–5.27) are independent of the initial conditions (5.8); this result holds generally, as will be shown in Sec. 6.

A simple approximate expression for W, the total equilibrium track population, may be obtained readily when $p\ll 1$. The quantities Δ_1 and Δ_2 are much larger than the remaining Δ_i, when $p\ll 1$, and further, the leading terms in D are

$$(1-p)^5\frac{1+pQ}{pQ}-(2-p)(1-p)\frac{1-p-pQ}{Q}.$$

Hence

$$W\doteq\frac{\Delta_1+\Delta_2}{D}$$
$$\doteq\frac{(2-p)Q}{1+pQ}\doteq 2Q. \tag{5.29}$$

For small association probability a track, once initiated, will be kept in the computer's memory the minimum period of time, 2 time units.

6. MARKOV CHAIN FORMULATION

A difference equation, or a system of difference equations, was used to describe birth and death processes in Secs. 4 and 5. It is possible to derive difference equations for the processes using Markov chains,[6,7] which will be done here. Aside from the insight which another treatment may provide, the Markov chain development enables one to unify the theory, and to make approximations readily.

Consider a given track at a fixed time. It may be considered to be in just one of a set of $M+1$ states $S_0, S_1, \cdots, S_M$. The subscript may denote m minus the number of successive nonassociations which the track has undergone, at the fixed time, if one is discussing birth and death Process A, or the various values of f, if Process B is being treated.

Let $P=\|p_{ij}\|$ be a matrix of transition probabilities, where p_{ij} is the (conditional) probability that if a track is in state S_i at time r, it will be in state S_j at time $r+1$. As examples one has: (a) birth and death—A

$$P=P_A=\begin{Vmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1-p & 0 & \cdots & 0 & p \\ 0 & 1-p & \cdots & 0 & p \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & 0 & p \\ 0 & 0 & \cdots & 1-p & p \end{Vmatrix}; \tag{6.1}$$

(b) birth and death—B

$$P=P_B=\begin{Vmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1-p & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & 0 & p \\ 0 & 0 & \cdots & 1-p & p \end{Vmatrix}, \tag{6.2}$$

[6] See J. L. Doob, *Stochastic Processes* (John Wiley and Sons, Inc., New York, 1953).

[7] See W. Feller, *An Introduction to Probability Theory and Its Applications* (John Wiley and Sons, Inc., New York, 1950), Vol. 1, for treatments of Markov processes.

and in particular, for the case treated before

$$P=\begin{Vmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1-p & 0 & 0 & p & 0 & 0 & 0 & 0 \\ 0 & 1-p & 0 & 0 & p & 0 & 0 & 0 \\ 0 & 0 & 1-p & 0 & 0 & p & 0 & 0 \\ 0 & 0 & 0 & 1-p & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 1-p & 0 & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 0 & 0 & 1-p & p \end{Vmatrix}. \quad (6.3)$$

The matrix P is, in every case, one which has just a single element, p_{00}, equal to one ($p<1$).

Let $p_{ij}^{(r)}$ be the probability that a track, starting in state S_i, will be in state S_j after exactly r time units.

One has¶

$$p_{ij}^{(r)}=\sum_s p_{is}p_{sj}^{(r-1)}=\sum_s p_{is}^{(r-1)}p_{sj}$$

or

$$\|p_{ij}^{(r)}\|=\|p_{ij}\|^r=P^r. \quad (6.4)$$

The formulation of birth and death processes in terms of Markov chains is now possible. Note that a track is "born" in a given state, S_I, and will not die unless it gets into the state S_0. One has, therefore, that the (conditional) probability of a track, born at time k remaining alive at time s, is given by

$$\sum_{j=1}^{M} p_{I,j}^{(s-k)}.$$

Let $N_b(i)$ be the average number of tracks born at time i. Then $\nu(r)$, the "expected" total track population, at time r will be given by

$$\nu(r)=\sum_{i=0}^{r} N_b(i)\Big(\sum_{j=1}^{M} p_{I,j}^{(r-i)}\Big),$$

with

$$p_{I,j}^{(0)}=1 \quad \text{if} \quad j=I, \quad \text{and} \quad p_{I,j}^{(0)}=0 \quad \text{if} \quad j\neq I. \quad (6.5)$$

Since

$$N_b(i)=QN[1-pn(i-1)], \quad i\geqslant 1 \quad (6.6)$$

one has

$$n(r)=Q\sum_{i=1}^{r}\sum_{j=1}^{M} p_{I,j}^{(r-i)}-Qp\sum_{i=1}^{r}\sum_{j=1}^{M} n(i-1)p_{I,j}^{(r-i)} +Q\sum_{j=1}^{M} p_{I,j}^{(r)}. \quad (6.7)$$

Furthermore, the $n(r;j)$ are given by the equations

$$n(r;j)=Q\sum_{i=1}^{r} p_{I,j}^{(r-i)}-Qp\sum_{i=1}^{r} n(i-1)p_{I,j}^{(r-i)} +Qp_{I,j}^{(r)}, \quad 1\leq j\leq M. \quad (6.8)$$

¶ See reference 6, p. 172.

Equation (6.7), or the system of Eqs. (6.8), is the desired formulation of birth and death.

The $p_{i,j}^{(r)}\to 0$ exponentially fast** as $r\to\infty$ ($i, j\neq 0$). Thus one may write, for $r\gg 1$, that

$$n(r)\doteq Q\Big[\sum_{k=1}^{s}\sum_{j=1}^{M} p_{I,j}^{(k)} -p\sum_{k=0}^{s}\sum_{j=1}^{M} n(r-k-1)p_{I,j}^{(k)}, \quad (6.9)$$

where the approximation has been carried to the sth order.

The solution of the difference equation (6.9) is

$$n(r)\doteq\mu_s+\sum_{k=0}^{s+1} c_k\xi_k{}^r\sim\mu_s. \quad (6.10)$$

The equilibrium population, if it exists, will be given by

$$n(\infty)\doteq\mu_s=\frac{Q\Gamma_s}{1+pQ\Gamma_s} \quad (6.11)$$

where

$$\Gamma_s=\sum_{k=0}^{s}\sum_{j=1}^{M} p_{I,j}^{(k)}.$$

Similarly, the detailed age distribution at equilibrium may be obtained from Eqs. (6.8) and (6.11).

One has, for large r,

$$n(r;j)\doteq Q\sum_{i=1}^{r} p_{I,j}^{(r-i)}-pQ \times\sum_{i=1}^{r}\Big\{\mu_s+\sum_{t=1}^{s} c_t\xi_t{}^{r-i-1}\Big\}p_{I,j}^{(r-i)}+Qp_{I,j}^{(r)} \quad (6.12)$$

or

$$n(\infty;j)\doteq Q\sum_{k=0}^{s} p_{I,j}^{(k)} -pQ\mu_s\sum_{k=0}^{s} p_{I,j}^{(k)}, \quad 1\leq j\leq M.$$

The w_i, given by Eqs. (5.20–5.28), were independent of the initial conditions. The generalization holds for other processes, as can be seen from Eq. (6.11); the initial conditions do not appear in that equation.

A numerical computation of the equilibrium population has been performed using $Q=\frac{1}{2}$, $p=\frac{1}{16}$, various powers of the matrix (6.3), and Eq. (6.11). It was found that μ_2 was in error by about 20 percent, μ_3 by about 12 percent, and μ_5 by about 8 percent. Thus the approximation is fairly accurate, in this case, even for small s.

** See reference 6, p. 179.

7. SIGNAL-TO-NOISE IMPROVEMENT

The purpose of the computations given in previous sections was to determine the total number of tracks, and hence the necessary memory capacity. It was asserted, in the Introduction, that the processes of pattern formation, maintenance, and rejection might serve to discriminate between signals and noise. The object of the present section is to determine quantitatively the signal-to-noise improvement.

Let $\nu_2'(r)$ be the number of signal tracks which are accepted as being real tracks at time r, $\nu_1'(r)$ be the number of noise tracks considered to be real tracks at time r, N_2 be the number of signal reports per frame time, and N_1 the number of noise reports per frame time. (The subscript 2 will refer to a signal, 1 to noise.)

The signal-to-noise improvement factor, I, may be defined by

$$I=\lim_{r\to\infty}\frac{\nu_2'(r)/\nu_1'(r)}{N_2/N_1}-1, \tag{7.1}$$

or

$$I=\lim_{r\to\infty}\frac{n_2'(r)}{n_1'(r)}-1, \tag{7.2}$$

if one uses the reduced variable $n(r)$. The primes denote only those tracks which are accepted as real.

(a) Process A

Here all tracks which have not suffered m successive nonassociations are considered to be real. One has, therefore, using Eq. (4.10),

$$I=\frac{Q_2}{Q_1}\cdot\frac{p_1}{p_2}\cdot\frac{1-(1-p_2)^m}{1-(1-p_1)^m}\cdot\frac{Q_1+(1-Q_1)(1-p_1)^m}{Q_2+(1-Q_2)(1-p_2)^m}-1. \tag{7.3}$$

When both $mp_1\ll 1$ and $mp_2\ll 1$, Eq. (7.3) becomes

$$I\doteq\frac{Q_2}{Q_1}\frac{1-mp_1(1-Q_1)}{1-mp_2(1-Q_2)}-1. \tag{7.4}$$

Using the numbers $Q_2=\frac{1}{2}$, $Q_1=\frac{1}{8}$, $p_2=\frac{1}{16}$, $p_1=\frac{1}{64}$, $m=3$, then $I\doteq 3.2$, from Eq. (7.4).

(b) Process B

All tracks whose firmness satisfies the condition $f_A\leq f\leq f_M$ are considered real. One may write

$$I=\lim_{r\to\infty}\frac{\sum_{f=f_A}^{f_M} n_2(r;f)}{\sum_{f=f_A}^{f_M} n_1(r;f)}-1, \tag{7.5}$$

or, using the notation $\lim_{r\to\infty} n_2(r;f)={}_2w_f$, $\lim_{r\to\infty} n_1(r;f)={}_1w_f$, as

$$I=\Big(\sum_{f=f_A}^{f_M}{}_2w_f\Big/\sum_{f=f_A}^{f_M}{}_1w_f\Big)-1. \tag{7.6}$$

In the particular case treated previously, $f_A=5$, $f_M=7$, and one has, using Eq. (5.20), and Eqs. (5.25–5.28),

$$I=\Big(\sum_{f=5}^{7}{}_2\Delta_f\Big/\sum_{f=5}^{7}{}_1\Delta_f\Big)\cdot\frac{D_1}{D_2}-1, \tag{7.7}$$

which becomes, when both $p_1\ll 1$ and $p_2\ll 1$,

$$I\doteq\frac{Q_2}{Q_1}\left(\frac{1-p_1}{1-p_2}\right)^3-1. \tag{7.8}$$

If the numbers $Q_2=\frac{1}{2}$, $Q_1=\frac{1}{8}$, $p_2=\frac{1}{16}$, $p_1=\frac{1}{64}$ are used again, then $I\doteq 3.6$, using Eq. (7.8). The signal-to-noise improvement is about the same for the two processes, with the above parameters and the same number of "real" states.

It is noteworthy that the signal-to-noise improvement factors are greater than 1, even though the equilibrium populations $n_2(\infty)$ and $n_1(\infty)$ are of the same order of magnitude as the number of signal reports and noise reports, per frame time, respectively. It is possible, therefore, to effect an improvement in signal-to-noise without imposing excessive memory requirements.

8. SUMMARY AND CONCLUSIONS

Two quantities of interest, the equilibrium population, and the signal-to-noise improvement factor, have been obtained for some birth and death processes.

It is to be noted that the equilibrium population is of about the same order of magnitude as the number of reports per frame time, if the total number of states is less than 10 [see Eq. (4.12) and Fig. 1] and if, further, the probability of an association is small. The total memory requirements may therefore be kept to reasonably small values if an effective method is used for eliminating tracks.

Of the two Processes A and B, studied here in some detail, Process B appears to be preferable both because the equilibrium population is smaller (if $m>2$) and I is slightly larger, than for Process A.

It should be emphasized that one may well expect the equilibrium signal population, and I, to be larger, in many applications, than the values computed here (for given initial Q_2 and p_2'). The assumptions listed in Sec. 2 may be more realistic for noise than for signal in a number of physical situations, since the long term correlation of signal observations has been ignored.

The present formulation thus may provide fairly crude lower bounds, rather than precise values, for $n_2(\infty)$ and I.

The author wishes to thank Dr. R. M. Brown, Dr.

R. I. Hulsizer, Dr. A. T. Nordsieck, Dr. H. Quastler, and Dr. J. P. Ruina for many helpful comments and discussions, and Dr. J. P. Nash, of the University of Illinois Digital Computer Laboratory, for coding help.

APPENDIX

It was remarked in Sec. 2 that the usual tests employed in statistics were not designed for the case of continuing observations. It is possible, however, to apply known methods (Neymann-Pearson, Sequential Analysis) to selected portions of a track's development, so that a choice of m and of l, f_I, f_A, and f_M can be made. A variety of modifications are possible; no attempt has been made to find the "best" ones.

Let H_0 be the hypothesis that a track is noise, and H_1, the alternative hypothesis that it is signal. Let α be the probability of committing an error of the first kind (false alarm) by accepting H_1 when H_0 is in fact true, and let β be the probability of an error of the second kind (miss) which occurs by accepting H_0 when H_1 is true.

(a) Determination †† of m

The probability that a signal track is born and then dies after the next m time units, is given by $Q_2p_2'(1-p_2)^m$. A simple way of determining m is to let m be the smallest integer, m_0, which satisfies the condition

$$Q_2p_2'(1-p_2)^{m_0}\leq\beta, \tag{A1}$$

where β is the desired miss probability.

The false alarm probability will be given by

$$\alpha=Q_1p_1'[1-(1-p_1)^{m_0}]. \tag{A2}$$

A similar method for choosing m could be used, by fixing α; β would then be determined once m_0 was found.

Other criteria, which depend on the long term behavior of a track rather than the initial behavior, as given here, can be used but they are somewhat more complicated.

†† This is an adaptation of Neymann-Pearson methods.

(b) Determination‡‡ of l, f_I, f_A, and f_M

Let α and β (both less than $\frac{1}{2}$) be given.§§ Define A_1 and A_2 by means of the equations

$$A_1=\beta/(1-\alpha), \tag{A3}$$

$$A_2=(1-\beta)/\alpha. \tag{A4}$$

One may determine f_I from the condition that f_I be the smallest integer satisfying

$$f_I\geq\log\left(\frac{Q_2p_2'}{Q_1p_1'}\cdot\frac{1}{A_1}\right). \tag{A5}$$

Similarly f_A is to be the smallest integer satisfying

$$f_A\geq\log\frac{A_2}{A_1}=\log\frac{(1-\alpha)(1-\beta)}{\alpha\beta}, \tag{A6}$$

and l the smallest integer satisfying

$$l\geq\frac{\log p_2/p_1}{\log(1-p_1/1-p_2)}. \tag{A7}$$

An additional condition must be imposed before f_M can be determined. One may use a criterion similar to the one given in Sec. (a), above. A signal track with $f=f_M$ will be considered to be "tentative" again if it suffers f_M-f_A+1 successive nonassociations. If δ is assigned, then one may determine f_M from the condition that f_M be the smallest integer satisfying

$$(1-p_2)^{f_M-f_A+1}\leq\delta. \tag{A8}$$

‡‡ The methods developed by Wald, in reference 4, are used here.

§§ The assumption that α and β are both less than $\frac{1}{2}$ is made for convenience, and is not necessary. See Wald (reference 4) for further details.

Role of Positive Ions in High-Voltage Breakdown in Vacuum

H. C. Bourne, Jr., R. W. Cloud, and J. G. Trump
Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received July 16, 1954)

Measurements were made of the electron emission from targets of magnesium, aluminum, steel, copper, gold, and lead when bombarded by ions from hydrogen, helium, nitrogen, xenon, and mercury. The ions, whose energy was varied over the range from 10 to 140 kilovolts, were produced by ionization of the residual gas in the anode electrode and emerged through a small opening into an essentially uniform accelerating field. An initial rapid rise of electron emission with ion energy was followed in all cases by a slow linear increase. The emission ratio varied from 2 to 20 and was maximum for nitrogen ions on steel. The effect of extractive gradient at the bombarded metal surface was slight and no simple dependence on the nature of the ion or of the bombarded metal was observed. The significance of these measurements on the electron-ion interaction theory of high-voltage breakdown in vacuum is discussed.

INTRODUCTION

EXPLANATION of high-voltage breakdown between metallic electrodes in vacuum depends on an understanding of the fundamental processes that contribute to interchange of charged particles and radiation across the gap. One of the more important processes suggested by Van de Graaff[1] in his theory of vacuum breakdown is electron emission from the cathode due to positive ion bombardment. Jackson, Bruining, Chaffee, Allee, Hill *et al.*,[2] and others have studied this emission with ion energies of 10 kv to 500 kv and found that from 2 to 20 electrons are emitted for each impinging positive ion. These data were obtained under conditions of low-cathode gradient and under the varying situations normally associated with different experimental setups.

The data reported here constitute a study of three basic factors which might influence electron emission caused by positive ion bombardment. These are (1) the effect of high extractive cathode gradients, (2) the effect of different metallic targets, and (3) the effect of different positive ions, including some of the heavier ions which may be encountered in a vacuum gap.

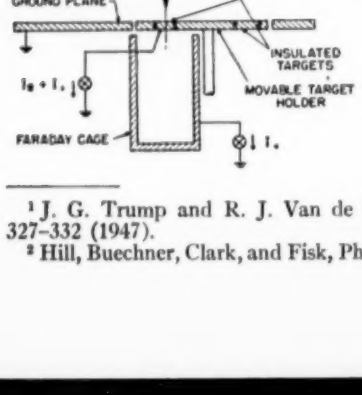


FIG. 1. Electrode arrangement for measuring the electron emission from several metals under high-energy positive-ion bombardment.

Under conditions of varying cathode gradient and total energy, targets of magnesium, aluminum, steel, copper, gold, and lead were bombarded by ions formed from helium, hydrogen, nitrogen (air), xenon, and mercury.

The electrode arrangement for measuring the secondary electron emission is shown in Fig. 1. Within the positive high-voltage terminal is the ion source, which includes a ring-shaped tungsten filament source of electrons and a cup-shaped anode. With an anode current of 2 milliamperes, an anode voltage of 1000 volts, and an ionization-gauge pressure of 6×10^{-5} mm of mercury, the collimated ion beam which emerged into the acceleration gap was about 0.07 microampere. These positive ions were characteristic of the residual gas in the vacuum chamber. To obtain the desired ion the corresponding gas was injected into the system through a leak at differential pressure. An appropriate concentration determined by an ionization-gauge reading was held in equilibrium with the vacuum pumps at a pressure well within the region of vacuum insulation. The lower electrode at ground potential contained an off-center disk, capable of being rotated, in which were alternatively a 0.5-inch hole and three 0.5-inch insulated targets of different metals. The Faraday cage on the center line of the system collected the positive-ion current when the ½-inch hole was centered. This

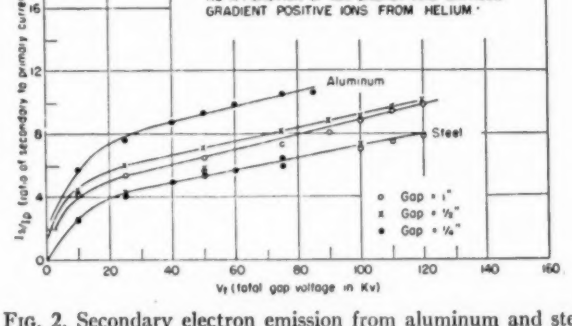


FIG. 2. Secondary electron emission from aluminum and steel under bombardment by helium ions with energies in the 10-kev to 120-kev range.

[1] J. G. Trump and R. J. Van de Graaff, J. Appl. Phys. **18**, 327–332 (1947).
[2] Hill, Buechner, Clark, and Fisk, Phys. Rev. **55**, 463–470 (1939).

Faraday cage was positively biased to prevent the extraction of secondary electrons from the cage by the electric field of the main gap.

With this arrangement the positive-ion current and its associated secondary electron emission could be determined. The ion-beam current was found directly as the cage current reading. The sum of the ion and secondary electron current was measured when the positive ion beam impinged on one of the insulated target elements. A cathode gradient, typical of high-voltage gaps in vacuum, was available at the target surface. This gradient and the total ion energy were varied by adjusting the terminal voltage and the gap length.

INFLUENCE OF CATHODE GRADIENT

Representative electron-emission curves for helium ions which illustrate also the influence of cathode gradient are shown in Figs. 2 and 3. The maximum

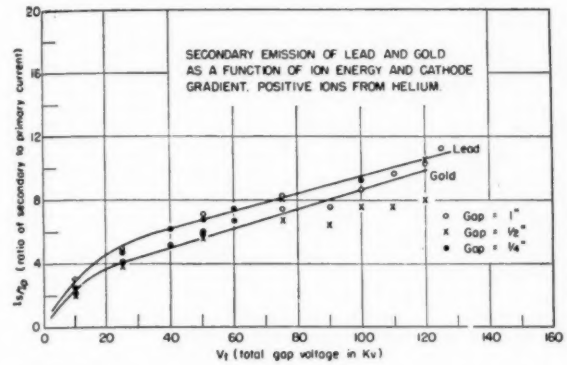

SECONDARY EMISSION OF LEAD AND GOLD AS A FUNCTION OF ION ENERGY AND CATHODE GRADIENT. POSITIVE IONS FROM HELIUM.
Lead
Gold
Gap = 1"
Gap = 1/2"
Gap = 1/4"
Is/Ip (ratio of secondary to primary current)
Vt (total gap voltage in Kv)

FIG. 3. Secondary electron emission from lead and gold under bombardment by helium ions with energies in the 10-kev to 120-kev range.

emission obtained from aluminum and steel bombarded by helium ions was about 11 electrons per positive ion. Each curve for aluminum represents a constant gap and hence varying energy and gradient. For aluminum and magnesium targets a small but definite increase of emission with cathode gradient was observed; this was slightly greater than could be accounted for by the various experimental variations. However, for steel and copper no dependence on cathode gradient was noticeable. In Fig. 3, which gives similar emission data for lead and gold, the same independence of electron emission on cathode gradient is observed. The other positive ions produced electron-emission trends consistent with those found with helium.

All of these emission curves have a shape characterized by an initially rapid rise followed by an almost linear increase at the higher energies. The data were recorded after equilibrium conditions had been established by an hour or more of positive ion bombardment. Although the system made use of Hg diffusion pumps with liquid-air trapping and the electrodes were

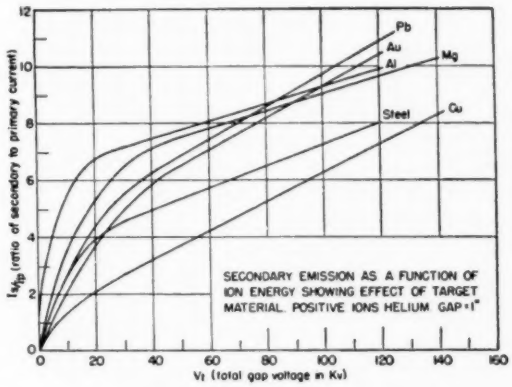

Pb
Au
Al
Mg
Steel
Cu
SECONDARY EMISSION AS A FUNCTION OF ION ENERGY SHOWING EFFECT OF TARGET MATERIAL. POSITIVE IONS HELIUM GAP=1"
Is/Ip (ratio of secondary to primary current)
Vt (total gap voltage in Kv)

FIG. 4. Secondary electron emission from several metals under bombardment by helium ions. The ion-energy range is 10 kev to 140 kev, and vacuum gap is 1 inch.

carefully prepared and cleaned, it is appreciated that the emission surfaces were complex and contaminated though representative of good vacuum conditions. An explanation of the increase of electron emission with ion energy can be found in Kapitza's theory[3] which recognized the transfer of the kinetic energy of the ion to electrons and their diffusion and escape from the bombarded surface. The penetration of the ions into the target material did not exceed the depth from which secondary electrons could escape.

EFFECT OF TARGET MATERIAL

Figure 4 shows the emission from different metallic targets bombarded by ions of helium. The targets were all buffed and polished carefully although the bombardment itself proved to be the most important conditioning agent. No correlation between the emission and the resistivity or heat conductivity of the metals seems to exist, but the secondary electron emission is least for targets of intermediate atomic number. In Fig. 5 the emission is plotted against the atomic number of the target material at two ion energies and with helium, nitrogen, and xenon ions. Curves of similar shape are obtained when emission is plotted against the metal density. Minimum electron emission is obtained from materials whose atomic number is about 40 and whose density is in the neighborhood of 9 grams per cubic

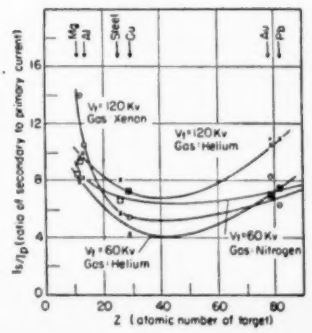


FIG. 5. Secondary electron emission under positive-ion bombardment plotted as a function of atomic number of the target material.

[3] P. L. Kapitza, Phil. Mag. **45**, 989–998 (1923).

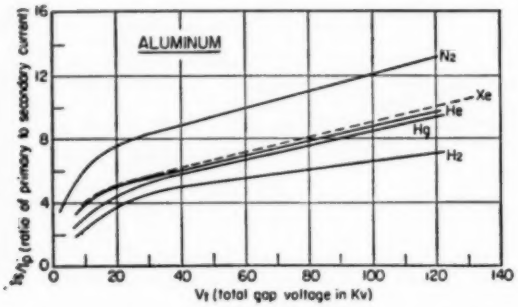


FIG. 6. Secondary emission from aluminum as a function of the energy of several types of positive ions.

centimeter. For metals of comparable density and atomic number this trend is not always consistent which suggests the influence of other factors.

EFFECT OF THE POSITIVE ION AND ITS ANGLE OF INCIDENCE

The effect on electron emission of bombardment by ions of different ionic mass was studied using ions of hydrogen, helium, nitrogen, xenon, and mercury. Figures 6 and 7 show the emission from aluminum and gold when bombarded by these five ion types. There is no large difference in the emission although the very light and the very heavy ions seem to produce less emission than nitrogen ions. These data are again presented in Fig. 8 with the emission plotted against the ionic weight. In these experiments the singly-charged atomic ion was assumed to dominate the ion-beam current.

Some experiments were carried out in which the angle of incidence of the ions was altered by machining slanting surfaces on the target disks. As expected the electron emission increased with angular departure from normal incidence, the emission being about doubled when the angle between ion beam and target surface was 30°.

POSITIVE IONS IN VACUUM BREAKDOWN

This investigation has further clarified previous evidence[4] that secondary emission of electrons by positive ion bombardment is not a decisive process in the particle-interchange theory of high-voltage breakdown in vacuum. In this chain-reaction theory of breakdown, the condition for instability is reached when the electrons produced by positive ion bombardment of the cathode can cause the release of the initial number of new positive ions at the anode. Beyond this limit the interchange of charged particles across the gap, each acquiring the full-gap voltage, increases exponentially. This initiating breakdown mechanism would naturally develop a variety of secondary effects including local heating and thermionic emission, leading ultimately to the passage of a spark or arc depending on the external circuit.

Since it has been established[1] that the positive ion emission from metals bombarded by electrons is generally less than one ion per thousand electrons (coefficient A), the electron emission measured in this work of less than twenty electrons per impinging positive ion (coefficient B) is seen to be quantitatively inadequate to produce the required condition for instability, $A \times B > 1$.

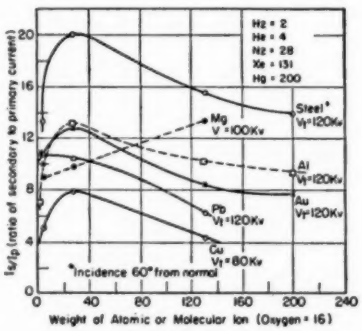


FIG. 8. Secondary electron emission as a function on the ionic weight of the bombarding ion.

No large increase in secondary electron emission was attained by using ions of higher atomic number and mass, or by increasing the electric field at the cathode surface. Secondary electron emission was moderately dependent on the nature of the bombarding particle, the nature of the bombarded metal and its contaminating surface layers, and on the angle of incidence. No factor capable of producing the substantial increase of electron-ion emission needed for a cumulative interchange of charged particles between the vacuum-immersed metals was observed. These experimental measurements did not, however, distinguish between electron and negative ion emission under positive ion bombardment. If the secondary emission included negative ions, these would be far more capable on returning to the anode with the full-gap energy to cause an adequate emission of positive ions. An analysis of the negatively charged emission is therefore planned.

The possibility also exists that vacuum breakdown is caused by larger atomic aggregates than the simple gaseous ions used in these experiments. Such aggregates, bombarding the cathode surface, would probably be

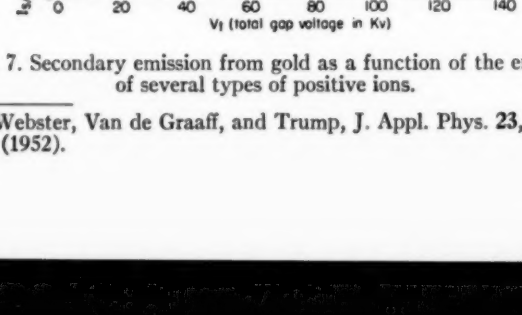


FIG. 7. Secondary emission from gold as a function of the energy of several types of positive ions.

[4] Webster, Van de Graaff, and Trump, J. Appl. Phys. **23**, 264–266 (1952).

multiply charged and would distribute their higher energy in a region more nearly at the surface and over a larger area. Cranberg[5] has suggested an ingenious "clump" theory of vacuum breakdown in which a charged clump of matter accelerated across the gap might precipitate breakdown by a single event at either electrode provided the total energy of the clump was above a critical amount. This theory of vacuum breakdown cannot be ruled out by these experiments even though the emission produced by the heavier positive ions did not show an important trend toward higher electron emission.

[5] Lawrence Cranberg, J. Appl. Phys. **23**, 518–522 (1952).

ACKNOWLEDGMENT

This work was supported in part by a grant to M.I.T. of the National Science Foundation. The advice and stimulation of Dr. R. J. Van de Graaff is gratefully acknowledged.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Nyquist's and Thevenin's Theorems Generalized for Nonreciprocal Linear Networks

R. Q. Twiss
Services Electronics Research Laboratory, Baldock, Herts, England
(Received July 8, 1954)

It is shown that an N-node nonreciprocal linear network, with a system of internal thermal noise generators at temperature T, is equivalent to the source-free network together with a set of correlated nodal current generators, with infinite internal impedance, such that $\langle i_r i_s^* \rangle_{Av} d\nu = 2kT(y_{rs}+y_{sr}^*)d\nu$ where y_{rs} is an element of the network admittance matrix and i_r is the rth nodal noise current with frequency ν. The noise sources can also be represented by a system of nodal voltage generators, with infinite internal admittance, such that $\langle v_r v_s^* \rangle_{Av} d\nu = 2kT(z_{rs}+z_{sr}^*)d\nu$.

The application of these results to a system of coupled antennas in an equi-thermal enclosure is noted and necessary conditions for the realizability of passive nonreciprocal networks are derived.

1. INTRODUCTION

RECENT developments in ferrites have made possible the design of nonreciprocal networks, embodying so called "gyrator" elements,[1] which have interesting applications to microwave rf circuits.[2,3] Because of internal losses, notably in the ferrite elements themselves, such a network will contain internal sources of thermal noise the exact location of which will not in general be known. The contribution of these thermal noise sources to the noise figure of a single receiver can be calculated by Nyquist's theorem. However if the "gyrator" is used, for example, to isolate one receiver from another, the finite loss in the ferrite elements implies that partially correlated thermal noise will be introduced into the two receivers. To calculate the magnitude of the correlation in this and similar cases one needs a generalized form of Nyquist's theorem which it is the main purpose of the present paper to provide.

A fundamental investigation of thermal noise in linear networks has been given by Schremp[4] and is freely made use of in the present discussion. The main results of this paper however are derived from first principles and cover the noise representations developed in the above reference,[4] which were given without proof, as special cases.

[1] B. D. H. Tellegren, Philips Research Repts. **3**, 81 (1948).
[2] C. L. Hogan, Bell System Tech. J. **31**, 1 (1952).
[3] J. H. Rowen, Bell System Tech. J. **32**, 1333 (1953).
[4] E. J. Schremp, M.I.T. Rad. Lab. Series **18**, Valley and Wallman, *Microwave Amplifiers* (McGraw-Hill Book Company, Inc., New York, 1949), Chap. 12.

2. NECESSARY CONDITIONS OF REALIZABILITY FOR THE NONRECIPROCAL NETWORK

Any one node of an $N+1$ node linear network can, without loss of generality, be regarded as grounded. The steady-state voltages and currents, at the N remaining nodes, of angular frequency ω, are then connected by the linear equations

$$V_r=\sum_{s=1}^{N} z_{rs}I_s, \quad r=1\cdots N, \tag{1}$$

$$I_r=\sum_{s=1}^{N} y_{rs}V_s, \quad r=1\cdots N, \tag{2}$$

where V_r is the voltage between the rth accessible node and earth, and $\sum_{r=1}^{N} I_r$ is the current flowing in the earth return.

If Z_{rs}, Y_{rs} are the co-factors of z_{rs}, y_{rs} in the expansion of the determinants

$$\Delta(z)=\|z_{rs}\|, \quad \Delta(y)=\|y_{rs}\|, \tag{3}$$

respectively, then it follows immediately that

$$y_{rs}=Z_{sr}/\Delta(z); \quad z_{rs}=Y_{sr}/\Delta(y) \tag{4}$$

and

$$\Delta(y)\cdot\Delta(z)=1. \tag{5}$$

It is no longer necessary that $y_{rs}=y_{sr}$ if the network is nonreciprocal but certain restraints must still be imposed upon the choice of the individual matrix elements if the network is to be physically realizable. In particular it is necessary that[4]

$$\begin{aligned} y_{rs}(-j\omega)&=y_{rs}^*(j\omega),\\ z_{rs}(-j\omega)&=z_{rs}^*(j\omega), \end{aligned} \tag{6}$$

while if the network is passive it is shown in Appendix I, by means of the first law of thermodynamics, that

$$y_{rr}+y_{rr}^*\geqslant 0, \tag{7}$$

$$\begin{aligned} [(y_{rs}+y_{rs}^*)+(y_{sr}+y_{sr}^*)]^2 \\ +|(y_{rs}-y_{rs}^*)+(y_{sr}-y_{sr}^*)|^2 \\ \leqslant 4(y_{rr}+y_{rr}^*)(y_{ss}+y_{ss}^*), \end{aligned} \tag{8}$$

together with similar constraints on the elements of the impedance matrix.

Similar, though weaker, constraints have been obtained by McMillan[5] for the special case of the three terminal network, and it is also possible to derive the inequalities directly from the noise theory of the present paper so that one might claim that they were consequences of the second law of thermodynamics as well as of the first law.

3. A GENERALIZED FORM OF THEVENIN'S THEOREM

The general linear network may contain a number of inaccessible nodes together with internal voltage and current sources the location of which may be unknown. However as long as there are only N independent accessible nodes such a system is indistinguishable from a source free network, with the same admittance or impedance matrix, together with a set of N nodal current generators, of infinite internal impedance, such that the current from the generator of the rth node is equal to the current flowing into the rth node of the original network when all the nodes of the latter are short-circuited to earth. The internal sources may alternatively be represented by a set of N nodal voltage generators of infinite internal admittance such that the voltage across the generator in the rth node is equal to the voltage across the rth node of the original network when all the nodes of the latter are open-circuited.

This generalization of Thevenin's theorem[6] holds true not only for coherent but also for thermal noise sources and can be proved by a straight forward extension of the familiar argument for the two terminal case.

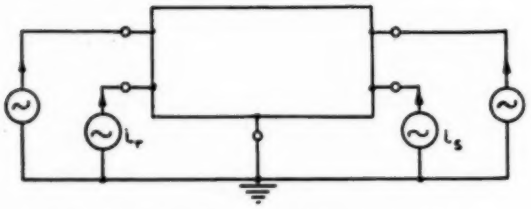


FIG. 1. Nodal current noise representation.

The equivalent nodal noise sources are not in general independent and we shall now find expressions for the average value of their cross correlations with the aid of the second law of thermodynamics.

4. A GENERALIZATION OF NYQUIST'S THEOREM

To begin let us consider the nodal current noise representation of Fig. 1. Following the conventional procedure we take a Fourier representation[4] for the nodal noise currents where $i_r(j\omega)$ represents the noise current at the rth node, with angular frequency ω, such that

$$\langle i_r(j\omega)i_r(-j\omega)\rangle_{\Delta\nu}$$

represents the average power dissipated by this noise current in unit resistance and unit band width.

Since the nodal noise currents are physically measurable quantities[4]

$$i_r(-j\omega)=i_r^*(j\omega)$$

and the expression for the average power dissipated by any linear combination of the N nodal noise currents consists of a sum of terms of the form

$$\langle i_r(j\omega)i_s^*(j\omega)\rangle_{\Delta\nu}.$$

The noise properties of the network are completely determined once such quantities are known for all r, s.

To calculate these correlations we assume that all accessible nodes, save the rth and sth, are shorted to ground and that the rth and sth nodes are grounded by an arbitrary resistive element, of conductivity G, and an arbitrary reactive element of susceptance jY, respectively. It then follows from the second law of thermodynamics that the average power dissipated in the network by the thermal noise currents associated with the resistive load must equal the average value of the power dissipated in the resistive load by the network noise currents, if both load and network are at the same temperature T, whatever the magnitude of G and Y; a result that is equally valid when the resistive and reactive loads are interchanged.

The noise current system associated with the network itself is shown in Fig. 2(a) and that associated with the resistive load in Fig. 2(b). For the case of Fig. 2(a)

$$\left.\begin{aligned} I_s&=i_s-jYV_s\\ I_r&=i_r-GV_r \end{aligned}\right\} \tag{9}$$

and since $V_t=0$ $t\neq r$, s we get from Eq. (2) that

$$\left.\begin{aligned} i_r-GV_r&=y_{rr}V_r+y_{rs}V_s\\ i_s-jYV_s&=y_{sr}V_r+y_{ss}V_s \end{aligned}\right\}, \tag{10}$$

so that

$$V_r=[(jY+y_{ss})i_r-y_{rs}i_s]/D, \tag{11}$$

[5] E. McMillan, J. Acoust. Soc. Am. **18**, 344 (1946).
[6] A. T. Starr, *Electric Circuits and Wave Filters* (Pitman, London, 1946), second edition, p. 78.

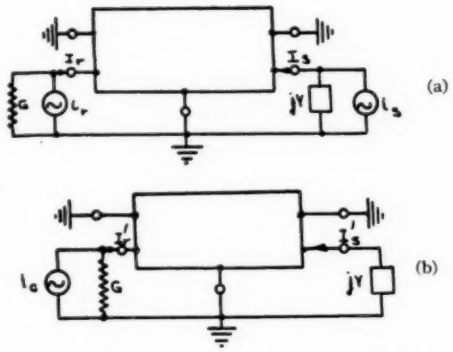


FIG. 2. (a) Network noise current system. (b) Load noise current system.

where

$$D=\begin{vmatrix} G+y_{rr} & y_{rs} \\ y_{sr} & jY+y_{ss} \end{vmatrix}. \tag{12}$$

The average power P_L absorbed in the load is proportional to $G\langle V_rV_r^*\rangle_{Av}$ so that

$$P_L=\frac{G}{DD^*}\{\langle i_ri_s^*\rangle_{Av}(y_{ss}+jY)(y_{ss}^*-jY) +\langle i_si_s^*\rangle_{Av}y_{rs}y_{rs}^*-\langle i_ri_s^*\rangle_{Av}(y_{ss}+jY)y_{rs}^* -\langle i_r^*i_s\rangle_{Av}(y_{ss}^*-jY)y_{rs}\}. \tag{13}$$

For the case of Fig. 2(b)

$$V_r'=(jY+y_{ss})i_G/D, \tag{14}$$

$$I_r'=i_G-GV_r'=[y_{rr}(y_{ss}+jY)-y_{sr}y_{rs}]i_G/D \tag{15}$$

and the average power P_N flowing into the network is proportional to $\frac{1}{2}\langle I_r'V_r'^*+I_r'^*V_r'\rangle_{Av}$ and is given by

$$P_N=\frac{\langle i_Gi_G^*\rangle_{Av}}{2DD^*}\{(y_{ss}^*-jY)[y_{rr}(y_{ss}+jY)-y_{sr}y_{rs}] +(y_{ss}+jY)[y_{rr}^*(y_{ss}^*-jY)-y_{sr}^*y_{rs}^*]\}. \tag{16}$$

Now from Nyquist's theorem[7] we know that

$$\langle i_Gi_G^*\rangle_{Av}d\nu=4kTGd\nu, \tag{17}$$

so that if $P_N=P_L$ for all Y we have, on equating coefficients of Y^2, that

$$\langle i_ri_r^*\rangle_{Av}d\nu=2kT(y_{rr}+y_{rr}^*)d\nu, \tag{18}$$

just as for the reciprocal network.

Interchanging the resistive and reactive loads, that is interchanging r and s we get

$$\langle i_si_s^*\rangle_{Av}d\nu=2kT(y_{ss}+y_{ss}^*)d\nu. \tag{19}$$

Equating the terms in P_N, P_L which are independent of and linearly proportional to Y, respectively, and eliminating $\langle i_ri_r^*\rangle_{Av}\langle i_si_s^*\rangle_{Av}$ by means of Eqs. (18) and (19), we get

$$\left.\begin{aligned} &y_{ss}y_{rs}^*\langle i_ri_s^*\rangle_{Av}+y_{ss}^*y_{rs}\langle i_r^*i_s\rangle_{Av} =2kT[y_{ss}y_{rs}^*(y_{rs}+y_{sr}^*)+y_{ss}^*y_{rs}(y_{rs}^*+y_{sr})] \\ &-y_{rs}^*\langle i_ri_s^*\rangle_{Av}+y_{rs}\langle i_r^*i_s\rangle_{Av} =2kT[-y_{rs}^*y_{sr}^*+y_{rs}y_{sr}] \end{aligned}\right\} \tag{20}$$

which may be solved to give

$$\langle i_ri_s^*\rangle_{Av}d\nu=2kT(y_{rs}+y_{sr}^*)d\nu, \tag{21}$$

$$\langle i_r^*i_s\rangle_{Av}d\nu=2kT(y_{rs}^*+y_{sr})d\nu. \tag{22}$$

To find the corresponding expressions for the correlations between the equivalent nodal voltage noise generators v_r we note that v_r is the open circuit noise voltage across the rth node, given by

$$v_r=\sum_{l=1}^{N}z_{rl}i_l, \tag{23}$$

where i_l is the l-node current generator.

Hence

$$\langle v_rv_s^*\rangle_{Av}d\nu=\sum_{l,m=1}^{N}z_{rl}z_{sm}^*\langle i_l\cdot i_m^*\rangle_{Av}d\nu. \tag{24}$$

Now from Eq. (21)

$$\langle i_li_m^*\rangle_{Av}d\nu=2kT(y_{lm}+y_{ml}^*)d\nu \tag{25}$$

while from Eq. (4)

$$y_{lm}=Z_{ml}/\Delta(z);\quad y_{ml}^*=Z_{ml}^*/\Delta^*(z). \tag{26}$$

But

$$z_{rl}Z_{ml}/\Delta(z)=\delta_{rm};\quad z_{sm}^*Z_{lm}^*/\Delta^*(z)=\delta_{ls}, \tag{27}$$

where δ_{rm} is the Kronecker delta, hence

$$\langle v_rv_s^*\rangle_{Av}d\nu=2kT(z_{rs}+z_{sr}^*)d\nu. \tag{28}$$

In the reciprocal network, where $y_{rs}=y_{sr}$

$$\langle i_ri_s^*\rangle_{Av}d\nu=\langle i_r^*i_s\rangle_{Av}d\nu=2kT(y_{rs}+y_{rs}^*)d\nu$$

a result already obtained[4] by Schremp,* and it is then possible to replace the system of correlated nodal noise currents by an equivalent system of uncorrelated intra-nodal currents. In the more general network however the lack of reciprocity destroys the simplicity of the intra-nodal representation which is no longer of practical interest.

Schremp[4] has discussed alternative noise representations appropriate to the case of inaccessible internal nodes and has stated the appropriate generalizations of Nyquist's theorem. A similar theory can be developed for the nonreciprocal case but it is too complicated and of too little practical value to give here.

The quantity $\langle(i_r-xe^{i\phi}i_s)(i_r^*-xe^{-i\phi}i_s^*)\rangle_{Av}$ is a positive definite quadratic form in x for all real ϕ. On substituting for $\langle i_si_s^*\rangle_{Av}$, etc. from Eq. (22), etc. and

[7] H. Nyquist, Phys. Rev. **32**, 110 (1928).

* In the present paper we have followed the usual procedure and considered only positive frequencies. Schremp's analysis however retains both positive and negative values of ω so his expressions for $\langle i_r(j\omega)i_s(-j\omega)\rangle_{Av}$ are smaller by a factor 2 than those given here.

applying the analysis given in Appendix I one obtains the inequalities (7), (8), which justifies the claim, made in Sec. 2, that these inequalities are consequences of the second law of thermodynamics.

5. DISCUSSION

We have shown that a general nonreciprocal network with a system of internal thermal generators all at temperature T is equivalent to the source-free network together with a system of correlated nodal noise current generators, with infinite internal impedance, such that

$$\langle i_r i_s^* \rangle_{Av} d\nu = 2kT(y_{rs}+y_{sr}^*)d\nu.$$

Alternatively the internal noise sources can be represented by a system of correlated nodal voltage generators, with zero internal impedance such that

$$\langle v_r v_s^* \rangle_{Av} d\nu = 2kT(z_{rs}+z_{sr}^*)d\nu.$$

In practice different parts of the network may be at different temperatures and the theory can then only be applied if the network can be decomposed into a set of physically separate networks each at a specific internal temperature. For example the system might consist of two coupled antennas connected into an rf gyrator system at room temperature T_2 which is in turn connected to two separate receivers the input impedance of which are at temperature T_3. If the antennas are in a blackbody enclosure at temperature T_1, the theory of this paper can be used to find the noise correlation between the two receiver channels, as long as the direct antenna coupling impedance at temperature T_1 is separated off from the circuit coupling at temperature T_2.

It follows immediately that the correlation between the open circuit antenna voltages is given by

$$\tfrac{1}{2}\langle (v_1 v_2^* + v_1^* v_2) \rangle_{Av} = 2kT_1(z_{12}+z_{12}^*),$$

where z_{12} is the mutual antenna impedance, and this is zero when the mutual coupling is purely reactive. This result seems to conflict with the conclusion of Mills and Little[8] that the correlation between the noise signals received by two antennas in a uniform temperature enclosure should be identically zero.

In the general case where the background temperature varies across the antenna polar diagrams, the simple theory of this paper breaks down and a detailed knowledge of the antenna polar diagrams is needed if the observed correlation is to be related to the actual distribution of background temperature.

[8] B. Y. Mills and A. G. Little, Australian J. Phys. 6, 272 (1953).

APPENDIX I

The power flow into the general linear network is

$$\sum_{r=1}^{N} \tfrac{1}{2}(I_r V_r^* + I_r^* V_r).$$

From Eq. (2) in the text this quantity may be written

$$\sum_{r,s=1}^{N} \tfrac{1}{2}(y_{sr}^* + y_{rs}) V_r V_s^*,$$

and, if the network is passive, this sum is to be positive for all possible values of V_r, V_s^*. If we put

$$V_r = x_r e^{-i\varphi_r} v_0, \tag{A.1}$$

where x_r is a real quantity, this condition means that the bi-quadratic form

$$\tfrac{1}{2}\sum_{r=1}^{N} \alpha_{rr} x_r^2 + \sum_{r\neq s=1}^{N} x_r x_s [\alpha_{rs}\cos(\varphi_r - \varphi_s) + \beta_{rs}\sin(\varphi_r - \varphi_s)],$$

where

$$\alpha_{rs} = y_{sr}^* + y_{rs} + y_{rs}^* + y_{sr},$$
$$\beta_{rs} = j[(y_{sr} - y_{sr}^*) - (y_{rs} - y_{rs}^*)] \tag{A 2}$$

must be positive definite for all φ_r, φ_s.

The necessary and sufficient conditions for this are that

$$\alpha_{rr} \geqslant 0,$$
$$[\alpha_{rs}\cos(\phi_r - \phi_s) + \beta_{rs}\sin(\phi_r - \phi_s)]^2 \leqslant \alpha_{rr}\cdot\alpha_{ss}. \tag{A.3}$$

The left-hand side of this latter inequality attains its maximum value of

$$\alpha_{rs}^2 + \beta_{rs}^2,$$

when

$$\tan(\phi_r - \phi_s) = \beta_{rs}/\alpha_{rs}$$

and on substituting for α_{rs}, β_{rs} in

$$\alpha_{r,r} \geqslant 0$$

and

$$\alpha_{rs}^2 + \beta_{rs}^2 \leqslant \alpha_{rr}\cdot\alpha_{ss} \tag{A.4}$$

from Eq. (A.3) we get the inequalities (7), (8) in the text.

Since the geometric mean of two quantities is always less than or equal to their arithmetic mean the inequality (8) is stronger than the second inequality derived by McMillan[5] for the special case of the dissipative three element network.

Impossible Behavior of Nonlinear Networks*

R. J. DUFFIN
Carnegie Institute of Technology, Department of Mathematics, Pittsburgh, Pennsylvania
(Received August 6, 1953; revised manuscript received July 8, 1954)

Networks of nonlinear inductors, nonlinear capacitors, nonlinear resistors, and direct current generators are considered. If the resistors are of quasi-linear type it is shown that such networks cannot generate alternating current or serve as power amplifiers.

THE aim of this note is to establish the following principle: *an isothermal electro-mechanical system whose primary resistors are quasi-linear cannot convert direct current to alternating current.*

Explanations of these terms are as follows:

The system is to be supplied with power by a set of direct current generators. These generators produce constant electromotive force. No other source of power—electrical or mechanical—is permitted. By "isothermal" it is meant that all parts of the system are maintained at essentially the same temperature by a suitable thermal bath.

The electrical part of the system is regarded as being composed of coils, capacitors, and resistors arbitrarily interconnected. The coils may be movable, and arbitrary magnetic coupling is permitted between them. This coupling may result from ferromagnetic material with arbitrary hysteresis. The capacitors may be variable and have a nonlinear dielectric material. The dielectric material may have hysteresis such as in "ferroelectrics."

The mechanical part of the system is coupled to the electrical part by magnetic and electrostatic forces. In particular, this could be by magnetostriction or electrostriction. The exact nature of the mechanical part of the system is of no concern for the problem being considered. As a special case the system may be entirely electrical and have no mechanical part.

A primary circuit is a simple closed circuit of the electrical network which contains a generator but no capacitors. A resistor is termed a *primary resistor* if it is in a primary circuit. Other resistors are termed *secondary*. Secondary resistors may be coupled to the generators but not by a direct current path.

A resistor is said to be *quasi-linear* if the voltage drop across the resistor is a function only of the current through the resistor, and if the differential resistance is positive. Of course a conductor which obeys Ohm's law is a quasi-linear resistor. Another example is a vacuum tube rectifier. A gas discharge tube exhibits hysteresis and so is not quasi-linear.

In this note switches and sliding contacts are regarded as variable resistors. These occur in relays and in commutators on motors. Such switches are not quasi-linear resistors. Electronic relays such as triodes and transistors are also not quasi-linear.

The system is said to "convert direct current to alternating current" provided there are sustained oscillations in some branches of the network. It is not supposed that these oscillations are necessarily sinusoidal. It is supposed that the system attains a *steady state* which continues for all time. At first this steady state will be taken to be a periodic state.

In other papers the writer has considered related questions involving the notion of quasi-linearity. The problem treated here is more general, but the conclusions are less specific.[1]

The electrical part of the system is diagrammed as a lumped network in a realistic fashion. Thus, an actual coil is represented as a coil with zero resistance in series with a resistor. The shunt and series resistance of an actual capacitor are separated in the diagram. The considerations that follow refer to this lumped network.

The network may be described as a set of junction points connected by branches. A branch is either: (1) a coil, (2) a capacitor, (3) a resistor, or (4) a generator. A direction is arbitrarily assigned to each branch. Relative to this direction let i denote the current through the branch and let v denote the potential drop (voltage) across the branch. In what follows $\sum$ denotes a summation over all branches, $\sum_1$ denotes a summation over coil branches, etc.

Consider an arbitrary set of branch currents satisfying Kirchhoff's first law and an arbitrary set of branch voltages satisfying Kirchhoff's second law. Then for such a virtual state of the network it is not difficult to show that

$$\sum vi=0. \tag{1}$$

If v and i correspond to an actual state of the network, then vi is the power absorbed by a branch. In this case (1) is a statement of the conservation of energy. If v_0 and i_0 correspond to another virtual state of the network, then

$$\sum (v-v_0)(i-i_0)=0 \tag{2}$$

because the sum or the difference of virtual states is a

* Presented at the Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, April 23–24, 1953. The work on this paper was sponsored by the Office of Ordnance Research, U. S. Army.

[1] R. J. Duffin, "Exponential decay in nonlinear networks," Proc. Am. Math. Soc. (to be published).

virtual state. For other applications of (1) see Bott and Duffin[2] and Tellegen.[3]

All actual resistors oppose the flow of current, so

$$vi \geq 0 \tag{3}$$

for a resistor branch. For a quasi-linear resistor it is assumed that v is a function of i such that $dv/di \geq \epsilon$ where ϵ is a positive constant. If i and i_0 are any two values of the current, it follows by the mean value theorem that $(v-v_0)/(i-i_0) \geq \epsilon$. Thus,

$$(v-v_0)(i-i_0) \geq \epsilon(i-i_0)^2. \tag{4}$$

Imagine that all the coils are short circuited and that all the capacitors are open circuited; there results a direct current network of resistors and generators. The only part of the network which is activated is the primary circuits. It is clear that this network has constant currents and voltages. (For a detailed analysis of this question, see Duffin.[4]) This is a virtual state of the original network. Denote it by v_0 and i_0. For the coil branches, $v_0=0$, and for the capacitor branches $i_0=0$. For a resistor which is not on a primary circuit, $v_0=0$ and $i_0=0$.

Let $W=-\int_0^T \sum_4 vi\,dt$, $Q_{12}=\int_0^T(\sum_1 vi+\sum_2 vi)dt$, and $Q_3=\int_0^T \sum_3 vi\,dt$. Here v and i denote a periodic state of period T. Then by (1)

$$W=Q_{12}+Q_3. \tag{5}$$

W is the work done on the system by the generators in a period. Q_3 is the heat supplied by the resistors to the thermal bath in a period. In principle the energy Q_3 does not have to be degraded but could be recovered as mechanical energy. To see this, imagine each resistor replaced by a direct current motor. The back emf of a motor would simulate the voltage-current relationship of a resistor. By the second law of thermodynamics it is impossible to derive energy from an isothermal cycle; so $W-Q_3 \geq 0$. Hence,

$$Q_{12} \geq 0. \tag{6}$$

Let v_0 and i_0 denote the virtual state with constant voltages and currents. Making use of relation (2) gives

$$\sum_1 v(i-i_0)+\sum_2(v-v_0)i+\sum_3(v-v_0)(i-i_0)=0.$$

The term $\sum_4$ vanishes because the generator voltages are the same in both states. Transposing gives

$$\sum_1 vi+\sum_2 vi+\sum_3(v-v_0)(i-i_0)=\sum_1 vi_0+\sum_2 v_0 i.$$

Integrating this relation over a period gives

$$Q_{12}+\int_0^T \sum_3(v-v_0)(i-i_0)dt = \sum_1 i_0\int_0^T v\,dt+\sum_2 v_0\int_0^T i\,dt. \tag{7}$$

For a capacitor branch $i=dq/dt$ where q is the charge on the capacitor. In the periodic state of the system q is periodic, so $\int_0^T i\,dt=q|_0^T=0$. Thus the terms of $\sum_2$ in (7) vanish.

Let the primary network be defined as the branches common to the primary circuits. Let F denote the magnetic flux linked with a circuit of the primary network. Let V denote the net voltage drop due to the coils in this circuit. Then by Faraday's law $V=dF/dt$. Since such a circuit contains no switches or sliding contacts, the quantity F is uniquely defined by the geometry and electrical state of the system. In the periodic state F is periodic,

$$\int_0^T V\,dt=F|_0^T=0. \tag{8}$$

For a coil branch of the primary network let $v'=\int_0^T v\,dt$ and for other branches let $v'=0$. It is a consequence of (8) that v' gives a virtual voltage state of the primary network. Since i_0 is a virtual current state of the primary network, it follows that $0=\sum v'i_0=\sum_1 v'i_0$. Thus the right side of (7) vanishes.

Write $\sum_3=\sum_p+\sum_s$ where $\sum_p$ is a summation over the primary resistors and $\sum_s$ is a summation over the secondary resistors. Then (7) becomes

$$Q_{12}+\int_0^T \sum_s vi\,dt+\int_0^T \sum_p(v-v_0)(i-i_0)dt=0.$$

By (3), (4), and (6) none of the terms in this relation can be negative; so they must all vanish identically. It follows from (4) that $i=i_0$ for a primary resistor. For a secondary resistor $iv=0$. It is recalled that the network is realistic; that is, all conductors are ascribed some resistance. Thus, $i=0$ for a secondary resistor. Hence all currents through the conductors of the network are constant. The only periodic state is a constant state; oscillations are impossible.

A steady state which is not periodic will at least be approximately periodic. Thus, if the system is in a given phase P at $t=0$, at later times the system must approximate the phase P with an arbitrary degree of precision. From a physical point of view there is no loss of generality in supposing that actually the system again takes on the phase P, say at $t=T$. It is clear that the same proof holds in this case.

It is permissible to suppose that some or all of the generators produce constant current rather than constant voltage.

The assumption that the system is isothermal may be omitted provided the system is such that all the heat produced in the resistors is degraded by thermal conduction to a constant temperature bath. The proof is unchanged.

The theorem stated has the following corollary: *an isothermal electro-mechanical system whose primary resistors are quasi-linear and which is powered by direct current generators cannot serve as a power amplifier.* The

[2] R. Bott and R. J. Duffin, Trans. Am. Math. Soc. **77**, 99–109 (1953).
[3] B. D. H. Tellegen, Philips Research Repts. **7**, 259–269 (1952).
[4] R. J. Duffin, Bull. Am. Math. Soc. **53**, 963–971 (1947).

justification of this statement depends on interpreting suitably the somewhat nebulous phrase "power amplifier." Examples of devices which can serve as power amplifiers are triodes, transistors, and relays. This is not the place to give a detailed analysis of the concept of a power amplifier. For the present purpose it is sufficient to postulate one of the characteristic properties of ordinary power amplifiers as follows: If a suitable feedback network is connected between the input and output terminals of a power amplifier, then steady oscillations result. Such a feedback network could be without direct current paths to the input and output terminals.

Now consider a system which satisfies the conditions of the corollary. The adjunction of a feedback network would result in a total system which satisfies the same conditions. By the theorem it follows that the total system could not generate sustained oscillations, and hence the original system could not have been a power amplifier.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Cross Correlation between Velocity and Current Fluctuations in Tube Noise*

S. V. YADAVALLI
General Electric Microwave Laboratory at Stanford, Palo Alto, California†
(Received July 8, 1954)

In practice one calculates the noise figure of an electron device by the Rack-Llewellyn-Peterson method. In this method one uses the fluctuations in current and velocity at the entrance plane as the input quantities. In a procedure such as the above, one needs to know the correlation between the velocity and current fluctuations at the entrance plane. It is shown in this paper that such a correlation between the above fluctuations is zero under certain conditions.

INTRODUCTION

JUST as in the case of a system the term "noise figure"[1] is used to signify the sensitivity or figure of merit of an electron tube (amplifier). As a fair approximation one calculates the noise figure of an electron tube by the successive use of the Llewellyn-Peterson equations.[2,3] This method of computation of noise in a tube is generally known as the Rack method[4] and also often called the Rack-Llewellyn-Peterson method. The above method has been successfully employed in the past by several workers[3,5,6] to study the noise properties of various devices and electron streams.

In the computation of noise figure of a tube or noise in an electron stream according to the Rack-Llewellyn-Peterson method, one needs to know the correlation between the velocity and current fluctuations. It is desired to discuss below such a correlation. Quite recently Watkins,[7] Pierce,[8] and also the author[9] employing different techniques have demonstrated that the assumption of lack of correlation between velocity and current fluctuations leads to a valid result in the problem of noise in a temperature limited electron beam in a drift space.

ANALYSIS

One can discuss the correlation between velocity and current fluctuations by using either (1) the cross-correlation function of the velocity and current fluctuations or (2) the cross-spectral density function of the velocity and current fluctuations.

The above two formulations should yield the same result since the cross correlation function and the cross-spectral density function are uniquely related through the Fourier transformation. We will use the first of the above two formulations below. Some details regarding the second approach are available elsewhere.[10]

Determination of Cross-Correlation Function of Velocity and Current Fluctuations

This section can be conveniently treated in two parts: (a) Rack's Noise Velocity, and (b) the cross-correlation function of velocity and current fluctuations. Although

* Part of this work was done at the Electronics Research Laboratory, University of California, Berkeley, California where it was supported by the U. S. Air Force, Wright Air Development Center, Dayton, Ohio under contract.

† Formerly Electronics Laboratory, General Electric Company, Syracuse, New York.

[1] H. T. Friis, Proc. Inst. Radio Engrs. **32**, 419 (1944).

[2] F. B. Llewellyn and L. C. Peterson, Proc. Inst. Radio Engrs. **32**, 144 (1944).

[3] L. C. Peterson, Proc. Inst. Radio Engrs. **35**, 1264 (1947).

[4] A. J. Rack, Bell System Tech. J. **17**, 592 (1938).

[5] D. A. Watkins, Tech. Rept. No. 31 (1951), Electronics Research Laboratory, Stanford University; L. D. Smullin, Tech. Rept. No. 142, (1949) Research Laboratory of Electronics, Massachusetts Institute of Technology.

[6] S. V. Yadavalli, J. Appl. Phys. **25** ,564 (1954).

[7] D. A. Watkins, J. Appl. Phys. **23**, 568 (1952).

[8] J. R. Pierce, Proc. Inst. Radio Engrs. **40**, 1675 (1952).

[9] S. V. Yadavalli, Quart. Appl. Math. **XII**, 105 (1954).

[10] S. V. Yadavalli, Interim Tech. Rept., Ser. No. 1, Issue No. 64, (1953), Electronics Research Laboratory, University of California, Berkeley.

the result of (a) is known in the past it is considered below for purposes of continuity of present argument.

(a) *Rack's Noise Velocity*

The analysis is carried out making the following assumptions; physical justification for these assumptions is given wherever possible.

(1) The fluctuation quantity associated with each electron is treated as being independent of the rest of the fluctuations. This allows us to neglect the correlation between the fluctuation quantities associated with two different electrons. This assumption may be justified on the basis that the electrons are emitted independently and randomly—a usual assumption in noise analysis. The rigorous justification of the above assumption can be attempted, however, only in the case of statistical mechanics[11] (assuming electrically noninteracting particles) where the correlation function of dynamical variables of two particles i and j is given by R_{ij}, and $R_{ij}\approx -1/n$, n being the number of particles in the system. Hence, in a case where n is large, we can assume that the correlation function is infinitesimally small.

(2) The fluctuation quantities are identically distributed. This is an assertion.

(3) The process is stationary in time[12] and observations are made after steady state is reached, that is

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T f(t)dt\to 0,$$

i.e.,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} f_j\xrightarrow[\text{a.s.}]{}0,$$

where f_j is the fluctuation quantity which is discrete and associated with the jth electron. It is tacitly assumed in the foregoing expressions that the limit exists. The justification for this is based on the argument that the infinite time average of a fluctuation is zero.

Given (1), (2), and (3) it follows from the theory of random processes that the expectation value or the average value of any fluctuation quantity f_j is zero.

Consider then at a plane infinitesimally close to the cathode[13] the velocities of electrons emitted from the cathode in a temperature limited case[14] (see Fig. 1). Let Δ_j be the probable interval of time in which the jth electron is emitted [or, more precisely, the interval of time between the emission of $(j-1)$th and $(j+1)$th electrons] from the cathode, and let the velocity of the jth electron be v_j. Note in the above that all Δ_j's (j is an index) are not equal but can be expressed in terms of a probability distribution (of Poisson type). Consider now the deviation δv_j of the instantaneous velocity of jth electron from the average velocity $\langle v\rangle$, given by $\delta v_j=v_j-\langle v\rangle\cdot\langle v\rangle$ is the infinite time average of the velocities of emission (see Fig. 1) or the average velocity as determined by a Maxwellian distribution. As stated above, by virtue of assumptions (1), (2), and (3) we have

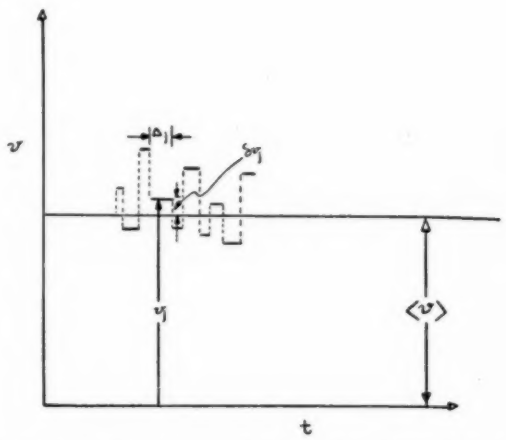


FIG. 1. Emission velocities of electrons at a plane infinitesimally close to the cathode.

$$\langle\delta v_j(t)\rangle=\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\delta v_j=0. \tag{1}$$

Also, by assumption (1)

$$\langle\delta v_j(t)\delta v_k(t+\tau)\rangle=\langle\delta v_j(t)\rangle\cdot\langle\delta v_k(t+\tau)\rangle=0,\quad \text{for}\quad \tau>\Delta_j. \tag{2}$$

Relation (2) may be called the cross-correlation function $R(\tau)$ of two pulses. Let us now proceed to consider a typical case.

Consider t in the foregoing relation (2) to lie in the interval $0\leqslant t\leqslant\Delta_j$; since each interval, Δ_j, shows the same behavior from a statistical point of view, we lose no generality in doing the following. If $(t+\tau)<\Delta_j$, that is $t<\Delta_j-\tau$, both quantities in Eq. (2) lie in the same interval Δ_j, and

$$\langle\delta v_j(t)\delta v_k(t+\tau)\rangle=(\delta v_j)^2,\quad \text{for}\quad (t+\tau)<\Delta_j. \tag{3}$$

Writing $R_{v_n}(\tau)$ for the auto-correlation function of fluctuations in velocities

$$R_{v_n}(\tau)=\lim_{T\to\infty}\frac{1}{T}\int_0^T\delta v(t)\delta v(t+\tau)dt. \tag{4}$$[15]

[11] A. I. Khinchin, *Mathematical Foundations of Statistical Mechanics* (translated from Russian by G. Gamow) (Dover Publications, New York, 1949).

[12] A random process characterized by certain probability distribution functions that are invariant under a change in the origin of time is called a stationary random process.

[13] We consider a plane close to the cathode so as to obtain the spectral density of velocity or current fluctuations. The transit time does not enter into such a discussion.

[14] The space-charge limited case will be treated at a later date.

[15] This integral is taken in the Lebesgue sense. Here the subscript is dropped to indicate that it is the over-all fluctuation that is used in the integral.

We also have for an interval Δ_j

$$\int_0^{\Delta_j} \delta v_j(t)\delta v_k(t+\tau)dt$$

$$= \int_0^{\Delta_j-\tau} (\delta v_j)^2 dt + \int_{\Delta_j-\tau}^{\Delta_j} 0\cdot dt$$

$$= (\delta v_j)^2(\Delta_j-\tau), \quad \text{for} \quad \tau<\Delta_j. \quad (5)$$

Since the time of observation (also time of integration) T is large, we can say that we have $T/\langle\Delta\rangle$ such intervals ($\langle\Delta\rangle$=average of Δ_j over all j) contributing the same amount. Hence, we can write

$$R_{v_n}(\tau) = \lim_{T\to\infty} \frac{\langle(\delta v_j)^2\rangle}{T}\frac{T}{\langle\Delta\rangle}(\langle\Delta\rangle-\tau).$$

That is,

$$R_{v_n}(\tau) = \begin{cases} \langle(\delta v_j)^2\rangle(1-\tau/\langle\Delta\rangle), & \text{for} \quad 0\leqslant\tau\leqslant\langle\Delta\rangle \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

We now apply the Wiener-Khinchin theorem which says that the auto-correlation function and the spectral density function are Fourier transforms of each other.[16] That is, if $G_{v_n}(f)$ is the spectral density in our case

$$G_{v_n}(f) = 4\langle(\delta v_j)^2\rangle\int_0^{\langle\Delta\rangle}(1-\tau/\langle\Delta\rangle)\cos 2\pi f\tau d\tau$$

$$= 2\langle\Delta\rangle\langle(\delta v_j)^2\rangle\left[\frac{\sin\pi f\langle\Delta\rangle}{\pi f\langle\Delta\rangle}\right]^2. \quad (7)$$

Note that $\langle\Delta\rangle=\nu^{-1}$, where ν is the average number of electrons emitted per unit time; hence, we can write

$$G_{v_n}(f) = \left[\frac{2(\langle v_j^2\rangle-\langle v_j\rangle^2)e}{I_0}\right]\cdot\left[\frac{\sin\pi f\langle\Delta\rangle}{\pi f\langle\Delta\rangle}\right]^2, \quad (8)$$

where, $I_0=\nu e$, e the charge of an electron and $\langle(\delta v_j)^2\rangle = \langle v_j^2\rangle-\langle v_j\rangle^2$. The value of $\langle(\delta v_j)^2\rangle$ in our present case is obtained by using a Maxwellian velocity distribution. For $f\langle\Delta\rangle\ll 1$, one obtains

$$G_{v_n}(f) \approx \left[\frac{2e}{I_0}(\langle v^2\rangle-\langle v\rangle^2)\right], \quad (9)$$

which is precisely the result obtained by Rack using a different procedure.

(b) *Cross-Correlation Function of Velocity and Current Fluctuations*[17]

Before proceeding to consider the cross correlation of velocity and current fluctuations the following comments are in order.

[16] S. O. Rice, Bell System Tech. J. **23**, 282 (1944).

[17] Again we consider a plane infinitesimally close to the cathode.

We consider the conduction current pulses due to individual electrons, and let the conduction current due to the jth electron[18] be i_j which does not exist at any time except in the interval Δ_j. Then the deviation δi_j of the instantaneous current of the jth electron from the average current I_0 is given by (i_j-I_0). I_0 is the infinite time average of current. Making again the same assumptions as in (a) it can be shown that one obtains the usual shot noise formula for $f\langle\Delta\rangle\ll 1$.

To discuss the cross correlation between the velocity and current fluctuations we proceed as follows.

Let $R_{i_nv_n}(\tau)$ be the cross-correlation function. Then from the definition of the cross-correlation function

$$R_{i_nv_n}(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \delta i(t)\delta v(t+\tau)dt, \quad (10)$$

using again in Eq. (10) the over-all fluctuation quantities. Note that $R_{i_nv_n}(\tau)=0$, for $\tau>\Delta$; since no correlation between individual electrons is assumed. By the same procedure as before and noting that $i_j=e/\Delta_j$, we obtain

$$R_{i_nv_n}(\tau) = \lim_{T\to\infty}\frac{e\langle(v_j-\langle v\rangle)\rangle}{T}\cdot\frac{T}{\langle\Delta\rangle}(\langle\Delta\rangle-\tau),$$

$$\text{for } 0\leqslant\tau\leqslant\langle\Delta\rangle, \quad (11)$$

$$\equiv 0,$$

since $\langle(v_j-\langle v\rangle)\rangle=0$, by virtue of assumptions (1)–(3).

We conclude from the above that the cross-correlation function of velocity and current fluctuations is zero under the conditions stated previously and this allows us to drop cross-power terms due to velocity and current fluctuations. However, this does not necessarily mean the velocity and current fluctuations are statistically independent, since vanishing of the cross-correlation function is *not* sufficient condition for statistical independence.[19]

STATISTICAL INDEPENDENCE OF VELOCITY AND CURRENT FLUCTUATIONS

Let us consider then the conditions under which the velocity and current fluctuations are statistically independent. First of all, note that if we have two random variables whose cross-correlation function is zero and which are Gaussian distributed, then the random variables are statistically independent. Also recall in our present problem the cross-correlation function is zero and the noise current amplitude can be described by a Gaussian distribution[16]; this, then suggests that we look into the possibility of replacing a Maxwellian velocity

[18] The cross correlation of velocity and current fluctuations has been discussed heuristically, based on the cross-spectral density function by G. Convert, Ann. radioélec. compagn. franc. assoc. T. S. F. **7**, No. 7, January (1952). This procedure is similar to the one given by the author in reference 10.

[19] H. Cramér, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, New Jersey, 1951).

distribution by a distribution which is Gaussian.[20] We also know that if a random variable is Gaussian distributed so also is the random variable minus a constant. Hence, the foregoing procedure would be equivalent to describing the velocity fluctuations by a Gaussian distribution.

The Maxwellian density function of velocities $f(v)$ can be written in the form

$$f(v)=\frac{mv}{kT_c}\exp-\frac{mv^2}{2kT_c},\quad 0\leqslant v<\infty,$$

$$\int_0^\infty f(v)dv=1,\quad\text{and}\quad\int_0^\infty v^n f(v)dv=\langle v^n\rangle,\quad n=1,2\cdots, \tag{12}$$

where m is the mass (of an electron in tube problems), k is the Boltzmann constant, T_c is the cathode temperature in degrees Kelvin absolute, and v is the velocity. A Gaussian distribution of velocity whose mean and mean square values are $\langle v\rangle$ and σ_v^2, respectively, can be written as

$$\varphi(v)=\frac{1}{(2\pi\sigma_v^2)^{\frac{1}{2}}}\exp\left(-\frac{(v-\langle v\rangle)^2}{2\sigma_v^2}\right),\quad -\infty<v<\infty. \tag{13}$$

We can then write the following expansion (see reference 13, page 222), sometimes called the Gram-Charlier series of type A:

$$f(v)=C_0\varphi(v)+\frac{C_1}{1!}\varphi'(v)+\frac{C_2}{2!}\varphi''(v)+\cdots+\frac{C_n}{n!}\varphi^{(n)}(v)+\cdots, \tag{14}$$

where the primes denote differentiation with respect to the argument, and

$$C_\nu=(-1)^\nu\int_{-\infty}^{\infty}H_\nu(\xi)f(\xi)d\xi. \tag{15}$$

H_ν is a Hermite polynominal of order ν.

$$C_0=1;\quad C_1=0;\quad C_2=0;\quad C_3=\frac{\mu_3}{\sigma_v^3};\quad C_4=\frac{\mu_4}{\sigma_v^4}-3;\quad\text{etc.},$$

the μ_n's are the central moments of order n; that is,

$$\mu_n=\int_0^\infty(v-\langle v\rangle)^n f(v)dv.$$

The series representation in Eq. (14) converges in the present case.[21] In view of the values of the C's just cited, let us see what happens if we neglect all higher order terms except the one of zero order. Note that this allows the emission velocities to be positive as well as negative. Since we do not have any negative emission velocities, we may terminate the Gaussian distribution (abruptly) at $v=0$. What does this mean? The suggested foregoing procedure (of abruptly terminating the Gaussian distribution at $v=0$) causes an error the upper bound of which can be estimated from the relation

$$\text{error}=\left\{1-\int_0^\infty\frac{1}{(2\pi\sigma_v^2)^{\frac{1}{2}}}\exp\left[-\frac{(v-\langle v\rangle)^2}{2\sigma_v^2}\right]dv\right\}. \tag{16}$$

Equation (16) tells us that by adopting the above procedure we are not considering about 2.8 percent of the total number of electrons. Note that we do not have to assume any cathode temperature for this calculation. It immediately follows then that within the framework of the approximation we can consider the velocity and current fluctuations to be statistically independent. Note, however, that the foregoing procedure suffers from the weakness that we do not know exactly the velocity range of the 2.8 percent electrons not considered.

Although we intend to come back to the space-charge limited case at a later date, the following comments are, however, in order. If one assumes that the conditions stated already (at the beginning of the paper) are satisfied at the potential minimum also, then the result of zero cross correlation between velocity and current fluctuations holds for the space-charge limited case as well. It is known, however, that under space-charge limited conditions the behavior of the region between the cathode and the potential minimum is nonlinear. Also, at the present time, no method is known for expressing the detailed behavior of the cathode to the potential minimum region in tractable mathematical terms. Consequently, the space-charge limited case can be treated only under some drastic assumptions which we do not intend to consider at the present time.

[20] The mean and mean square values are same with both distributions.

[21] The convergence of such a series can be investigated using a criterion given by Cramér, see p. 223, reference 13.

Variation of Magnetoresistivity of AuCu with Degree of Order

BERNARD WIENER, PHILIP SCHWED, AND GERHART GROETZINGER
Lewis Flight Propulsion Laboratory, National Advisory Committee for Aeronautics, Cleveland, Ohio
(Received July 8, 1954)

The room temperature resistivity and the magnetoresistivity of AuCu have been investigated as a function of the degree of order. Measurements were made first of all on samples quenched to room temperature after reaching equilibrium at temperatures below the Curie point. In addition measurements were made on samples which were first disordered by quenching from a temperature above the Curie point, were then annealed for various lengths of time at 265°C, and were finally quenched to room temperature. The coefficient of magnetoresistivity equals 1.8×10^{-12} and 6×10^{-14} gauss^{-2}, respectively, for the completely ordered and completely disordered alloy. The equilibrium curve for the magnetoresistivity shows a minimum at 365°C and a pronounced maximum at 388°C which is close to the transition temperature for the phase change of AuCu I to AuCu II. No such extremes in the value of the coefficient of magnetoresistivity occur during the annealing of a thermally disordered sample.

INTRODUCTION

IN the temperature range below 500°C the binary alloy AuCu undergoes two types of transformations[1] which in this case are closely related physically. In the first place it displays three crystal structures in this temperature range, being tetragonal (AuCu I) below 380°C, orthorhomic (AuCu II) between this temperature, and 408°C and face-centered cubic (AuCu III) above the latter temperature. In the second place it undergoes an order-disorder transformation for which the Curie point is at the temperature of the second transformation.

In conducting investigations of this alloy, it has often proved convenient to make use of the fact that at temperatures below 420°C these transformations proceed sufficiently slowly so that it is possible to freeze in the configuration corresponding to any such temperature by suitable quenching and thus to carry out the measurements at room temperature.[2] The alloy is strictly not in equilibrium in such a supercooled state; we shall, however, refer to supercooled states as equilibrium or nonequilibrium states depending on whether the alloy was in equilibrium before quenching or not. By using samples quenched from a disordered state it is possible to study the kinetics of the ordering and of the transformation of the crystal structure at any temperature below the Curie point. In particular, it appears that in AuCu during annealing at temperatures below 380°C the face-centered cubic structure changes directly into the tetragonal structure.[1]

One very useful approach which has been frequently applied to the study of the state of such alloys is represented by the determination of the electrical resistivity which gives an indication of the distribution of the atoms over the lattice (degree of order). For most alloys of this sort, the resistivity is considerably smaller in the ordered state than in the disordered state although in the case of CoPt[3] an increase in resistivity with increasing order has been observed. The decrease which is ordinarily observed is simply due to the decrease of the lattice scattering of the electrons carrying the electric current when a superlattice is established. The electronic states enter into the resistivity only through the number of free electrons, a quantity that does not change very much during the ordering process in AuCu.[4] However, due to the rearrangement of the lattice during ordering, the quantum states of the electrons near the Fermi surface[5] are changed as a consequence of the development of new Brillouin zone boundaries. Such changes will, in general, cause a considerable alteration in the values of the galvanomagnetic coefficients; and, thus, by appropriate measurements of these coefficients, it is possible to arrive at an estimate of this effect. Sidorov[6] determined the equilibrium curves of both the resistivity and the Hall coefficient of AuCu as a function of temperature. His results for the relation between resistivity and temperature are shown in Fig. 1. Of special interest is the fact that at the temperature at which the transformation AuCu I to AuCu II occurs, the resistivity increases quite rapidly with increasing temperature over a range of 10°C and then remains substantially constant almost up to the Curie point. The curve for the absolute value of the Hall coefficient (not shown in Fig. 1) has a similar shape.

In the case of AuCu the Hall coefficient is always negative in sign; whereas in the case of $AuCu_3$, it is negative when the alloy is in the disordered state and becomes positive when the alloy is ordered. From these results Sidorov concluded that in AuCu the states of the electrons near the Fermi surface were not very much affected by the transition. On the other hand

[1] C. H. Johannson and J. O. Linde, Ann. Physik **25**, 1 (1936).
[2] Källbäck, Nyström, and Borelius, Ing. Vetenskaps Akad. Handl. **157**, 3 (1941).
[3] Newkirk, Smoluchowski, Geisler, and Martin, J. Appl. Phys. **22**, 290 (1951).
[4] C. K. Sidorov, J. Exptl. Theoret. Phys. U.S.S.R. **16**, 503 (1946).
[5] T. Muto, Sci. Repts. Inst. Phys. Chem. Research, Tokyo **34**, 741 (1938).
[6] C. K. Sidorov, Izvest. Akad. Nauk. S.S.S.R., Ser. Fiz. **11**, 511 (1947).

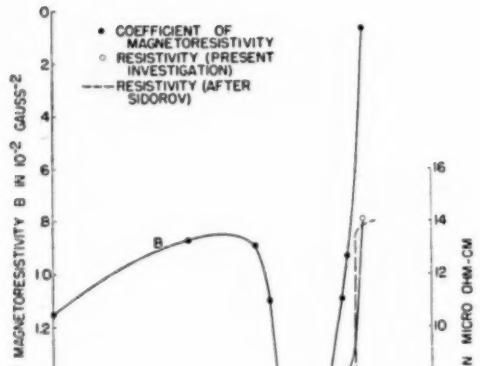


FIG. 1. Room temperature values of the resistivity ρ and the coefficient of magnetoresistivity B of AuCu as a function of temperature at which equilibrium was established.

Nicholas[7] has concluded, on theoretical grounds, that the energy difference between the ordered and disordered states in AuCu is largely to be attributed to the difference in the energies of the electrons near the Fermi surface in the two states.

According to the electron theory of metals, a sensitive measure of such changes of the electronic states as reflected by the resulting distortion of the Fermi surface in $\mathbf{k}$-space can be obtained from simultaneous measurements of the magnetoresistivity and the resistivity. We, therefore, decided to measure these two quantities as a function of temperature for samples in the equilibrium state in order to gain a better insight into the change of electronic states during the order-disorder transition and the accompanying phase transformations. In addition, we made similar measurements to study the changes occurring during the annealing of a sample completely disordered by quenching from a temperature above the Curie point.

To explain the relationship between the electronic states of the alloy and the electrical properties mentioned, we summarize some of the results of the electron theory of metals.[8] We will assume that we are dealing with a material composed of a large number of randomly oriented single crystals. Then the expression for the resistivity is given by

$$\rho = m/(e^2 n \bar{\tau}), \tag{1}$$

where e and m have the conventional meanings, n is the effective number of free electrons per unit volume, and $\bar{\tau}$ is the mean value of the relaxation time for electron scattering by the lattice averaged over the Fermi surface in $\mathbf{k}$-space. As mentioned earlier, it appears that, in AuCu, n remains substantially constant during the transformation while $\bar{\tau}$ increases with increasing degree of order. Thus, for this alloy ρ may be considered a measure of $1/(\bar{\tau})$.

The expression for the magnetoresistivity is more complex. For a magnetic field H, satisfying

$$H \ll mc/e\bar{\tau} \tag{2}$$

such as we employed in our work, the relative change in resistivity $\Delta\rho/\rho$ due to the application of a transverse magnetic field is represented by an expression of the form

$$\Delta\rho/\rho = BH^2 \tag{3}$$

with B a constant. The coefficient of magnetoresistivity, B, is given by

$$B = (e/mc)^2 \frac{(f^2\tau^3)_{Av}\bar{\tau} - (f\tau^2)_{Av}^2}{(\bar{\tau})^2}, \tag{4}$$

where $f(\mathbf{k})$ is defined as the ratio of m to $m^*(\mathbf{k})$, $m^*(\mathbf{k})$ being the effective mass of an electron with propagation vector $\mathbf{k}$; and where the averages, as before, are taken over the Fermi surface in $\mathbf{k}$-space. It is evident that the magnitude of B is governed by the value of $\bar{\tau}$ and by the extent of the fluctuations of τ and f over the Fermi surface, vanishing if the latter two quantities are constant. In order to obtain a measure of these fluctuations which is independent of changes in $\bar{\tau}$ during the order-disorder transition, we introduce the reduced coefficient of magnetoresistivity q^2 which is defined as

$$q^2 = \rho^2 B = 1/(enc)^2 \frac{(f^2\tau^3)_{Av}\bar{\tau} - (f\tau^2)_{Av}^2}{\bar{\tau}^4} \tag{5}$$

and is a measure of the degree of variation of f and τ if we assume n fixed. We shall find it convenient to express our experimental results in terms of this quantity. In interpreting these results it is useful to recall that any asymmetry appearing in τ is supposed to be a consequence of the fluctuation of f,[8] so that q^2 represents a direct measure of the distortion of the Fermi surface.

EXPERIMENTAL PROCEDURE

The samples used in this investigation were AuCu wires* 0.0635 cm in diameter and 10 cm in length which a chemical analysis showed to contain 50.2-atomic percent gold and 49.8-atomic percent copper.† All our samples were first heat treated in a vacuum furnace at a temperature of 750°C for two hours and then quenched by very rapid immersion in water. The

[7] J. P. Nicholas, Proc. Roy. Soc. (London) **66**, 201 (1953).
[8] See for example, H. Froehlich, *Elektronentheorie der Metalle* (Verlag Julius Springer, Berlin, 1936).

* The alloy was prepared by Horizons Inc. of Cleveland, Ohio.
† We are indebted to Mr. Dumas Otterson of this Laboratory for the analysis.

thin oxide film, which was formed during the quenching operation, was removed with 20 percent HCl. After this treatment the resistivity of the samples was found to be 14.1 μohm cm and the coefficient of magnetoresistivity, 6.0×10^{-14} gauss^{-2}.

To obtain the equilibrium curve we proceeded as follows. Some of the wires which had been disordered as described were partly ordered by being kept at a temperature of 325°C for 12 hr after which time they had a room temperature resistivity of 5.4 μohm-cm. One ordered and one disordered sample were brought to each desired (equilibrium) temperature and quenched after a suitable length of time. The periods used to bring the samples to equilibrium were selected in accordance with the results of Källbäck, Nyström, and Borelius[2] on the speed of this process and by the requirement that the final resistivities of the initially "ordered" and initially disordered samples agree within 2 percent. For the study of the annealing process a group of 15 disordered wires were annealed at 265°C for various lengths of times in a salt bath (for the shorter times) or in a vacuum furnace (for periods exceeding 30 min where thermal lags were less important) and then quenched by rapid immersion in water.

In the determinations of resistivity and magentoresistivity a Kelvin bridge was used for the necessary resistance measurements. The total current supplied to the bridge, this current being very close to that through the sample, was fixed at 2.25 amp by the use of a combination of Amperite ballast tubes. During the measurements, the sample was held firmly between the poles of an electromagnet in a Bakelite mount which was immersed in a constant temperature kerosene bath kept at 20°C. The pole pieces of the magnet were shaped so as to give a uniform magnetic field over the part of the sample between the two voltage electrodes (spaced 5.4 cm apart). In order to keep the field constant, a regulated current supply was used to excite the magnet. The resistivity was simply obtained by making a resistance measurement with the magnet unexcited. Each value of the magnetoresistivity represents the average of ten measurements carried out as follows. For five of these, the resistance was determined alternately with the sample in zero field and in a field of 12 000 gauss in one direction; and for the other five the same procedure was followed except that the direction of the field was reversed.

RESULTS AND DISCUSSION

The lower curves of Fig. 1 show the room temperature resistivity of AuCu as a function of the temperature at which the alloy was allowed to come to equilibrium prior to quenching as obtained by Sidorov and by ourselves. The upper curve of the same figure represents our results for the coefficient of the magnetoresistivity plotted on a reversed scale to simplify comparison with the resistivity curves. The magnetoresistivity curve displays two features of especial interest. In the first place, B is much larger for a completely ordered sample than for a completely disordered one; specifically it equals 1.8×10^{-12} gauss^{-2} and 6×10^{-14} gauss^{-2}, respectively, for the two states. In the second place, the magnetoresistivity does not decrease steadily with increasing temperature, but shows a well-defined minimum at a temperature of 365°C and a strong maximum of 1.9×10^{-12} gauss^{-2} at 388°C. It can be seen that these temperatures locate approximately the beginning and the end of the region of rapid increase of resistivity occurring near the temperature of the AuCu I→AuCu II transition.[1]

The curves of Fig. 2 represent our results for the resistivity and the coefficient of magnetoresistivity of initially disordered samples of AuCu as a function of the logarithm of the time of annealing at 265°C with the ordinates arranged as in Fig. 1. It can be seen that the maximum spread in the measurements of wires which have all been annealed the same length of time amounts to about 0.5 μohm-cm for ρ and to about 5×10^{-14} gauss^{-2} for B. These differences presumably result from slight variations in the states of individual wires as quenched. As is usually observed in such experiments, the resistivity drops quite swiftly at the beginning, but the rate of change diminishes in magnitude very markedly for long times. In general, this curve is in good agreement with those obtained by Stello[9] and by Schwed and co-workers[10] in similar experiments. The annealing curve for B is similar to that for ρ, in the sense that both quantities change

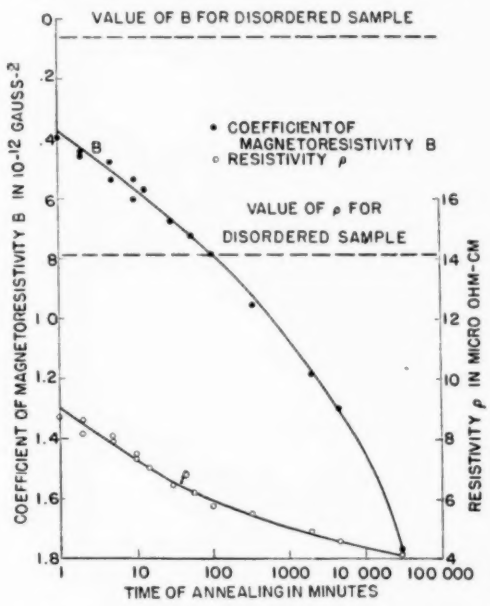


FIG. 2. Room temperature values of the resistivity ρ and of the coefficient of magnetoresistivity B of thermally disordered AuCu as a function of the time of annealing at 265°C.

[9] C. G. Stello, Report N.A.A., S.R. 171, North American Aviation, Inc., April, 1952.

[10] Schwed, Cooper, and Webeler (private communication).

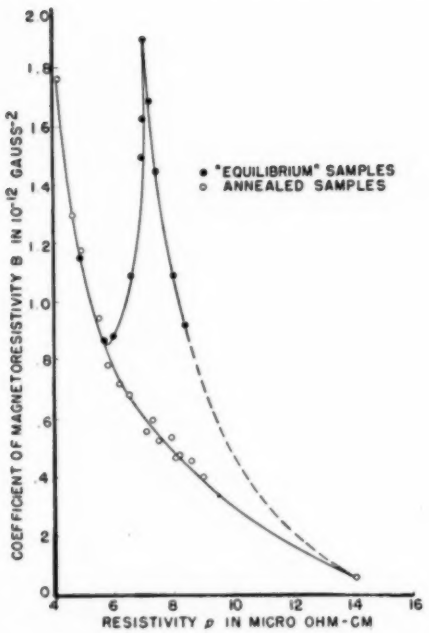

COEFFICIENT OF MAGNETORESISTIVITY B IN 10⁻¹² GAUSS⁻²
RESISTIVITY ρ IN MICRO OHM-CM
• "EQUILIBRIUM" SAMPLES
○ ANNEALED SAMPLES

FIG. 3. Coefficient of magnetoresistivity B of AuCu as a function of the room temperature resistivity ρ for samples in "equilibrium" and for thermally disordered samples annealed at 265°C.

monotonically with time and that the absolute values of the rates show the same trends.

For the sake of a direct comparison of the results of the equilibrium and annealing measurements, the coefficient B and the reduced coefficient q^2 of the magnetoresistivity are plotted in Fig. 3 and Fig. 4, respectively, as functions of the resistivity ρ. It will be noticed that as might be expected from Figs. 1 and 2, the equilibrium curves are quite different from the annealing curves. For the equilibrium samples the curve of q^2 against ρ displays essentially the same features as that of B against ρ. The quantity q^2 is much larger in the ordered state than in the completely disordered one, the respective values being 29.7 and 11.8×10^{-24} (ohm cm)^2 gauss^{-2}. The B *versus* ρ curves show a minimum at 5.7 μohm-cm and a maximum at 7.0 μohm-cm; extremes occur at nearly the same values for the q^2 against the ρ curve. In the case of the annealed samples, however, the shape of the q^2 against the ρ curve differs from that of the B against the ρ curve insofar as the former displays a minimum at 5.2 μohm-cm and a maximum at 8.2 μohm-cm while neither extreme appears in the latter curve. It should be noted that the positions of these extremes very nearly agree with those for the equilibrium samples.

The fact that q^2 is considerably larger when the alloy is in the ordered state then when it is in the completely disordered state implies that the Fermi surface is much more distorted in the former case than in the latter and that the states of the electrons with energies close to the Fermi surface are considerably altered during the order-disorder transformation. Furthermore, it would appear from the rapid increase of q^2 with temperature near 380°C for the equilibrium curve that the Fermi surface departs more radically from sphericity when the crystal structure is orthorhombic than when it is tetragonal. The subsequent decrease with rising temperature would then be a consequence of the continued diminution of the degree of order as the alloy approaches the Curie temperature. The fact that in the case of the annealing curves no maximum or minimum occurs in B and only relatively much smaller extremes in q^2 is in general agreement with the result of Johannson and Linde[1] that the alloy goes directly from the face-centered cubic into the

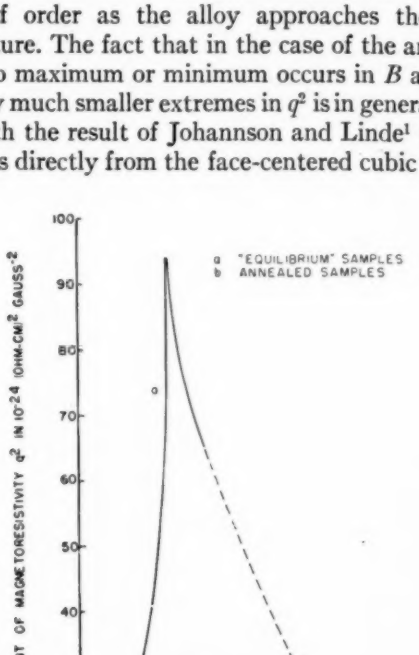

REDUCED COEFFICIENT OF MAGNETORESISTIVITY q² IN 10⁻²⁴ (OHM-CM)² GAUSS⁻²
RESISTIVITY ρ IN MICRO OHM-CM
a "EQUILIBRIUM" SAMPLES
b ANNEALED SAMPLES

FIG. 4. Reduced coefficient of magnetoresistivity q^2 of AuCu as a function of the resistivity ρ for samples in equilibrium and for thermally disordered samples annealed at 265°C. (Derived from the curves of Fig. 3.)

tetragonal phase during annealing. However, the presence of the extremes in the curve q^2 against ρ does indicate that a small admixture of alloy in the orthorhombic state may be formed in the course of the annealing at 265°C.

In conclusion, it may be remarked that the ratio of the value of q^2 in the ordered to the value of q^2 in the disordered state for AuCu is about $2\frac{1}{2}$ while according to Sidorov this ratio is about 10 for $AuCu_3$. It follows that, although in AuCu the Hall coefficient fails to change sign during the order-disorder transformation, nevertheless a considerable distortion of the Fermi surface and, hence, a considerable change of the states of the highest-energy electrons occurs also in this case.

Equivalent-Radius Method for Determination of Non-Newtonian Flow Curves from Viscometric Data*

K. H. CRAM† AND J. C. WHITWELL‡
Textile Research Institute and Department of Chemical Engineering, Princeton University, Princeton, New Jersey
(Received July 29, 1954)

A rapid and precise method is presented for approximating absolute flow curves (shear rate *vs* shear stress) of non-Newtonian liquids from data obtained with rotational or capillary viscometers. It is shown that the apparent Newtonian flow curve computed from a single measurement (i.e., a single flow rate) in either type of viscometer crosses the true flow curve at a shear stress corresponding to some intermediate radius between the bob and cup radii or less than the capillary radius. If shear rate and stress are computed at the common point using the ordinary Newtonian relations for shear stress and shear rate, the point so calculated will lie on the absolute flow curve. An "equivalent radius" rather than a wall radius is used in the conventional Newtonian equations.

The position of the equivalent radius is fairly insensitive to changes in the form of the flow curve, particularly in the case of the rotational viscometer. A method for determining the proper intermediate radius is developed and graphs are presented to facilitate this determination and permit estimation of the accuracy of the method. An example is given using data from the literature [I. M. Krieger and S. H. Maron, J. Appl. Phys. **25**, 72 (1954)] for a 62.2 percent GR-S Latex slurry. Within the range of the data (slope of the log-log flow curve varying from 1.5 to 3.0), the maximum error incurred by use of a single equivalent radius for each of two rotational viscometers was 0.1 percent for the instrument with a cup/bob radius ratio of 1.064, and 0.4 percent for the instrument with a ratio of 1.113. Accuracy can be increased by use of an iterative procedure.

The method has been developed in detail only for the case of pseudo-plastic flow, but the principle can be applied to other types of flow.

INTRODUCTION

ABSOLUTE flow curves of non-Newtonian liquids are generally determined from data obtained with capillary or rotational viscometers. This paper presents a method for deriving absolute flow curves from a series of measurements of flow rates and pressure gradients in a capillary viscometer or of rotational speeds and torques in a rotational viscometer. A rigorous method for the capillary-viscometer case was first developed by Weissenberg and Rabinowitsch.[1] They, and others,[2–4] developed a differential equation which can be solved fairly readily by graphical differentiation of the experimental data. The rigorous mathematical solution for the rotational-viscometer case yields a difference equation[3] which is not so readily susceptible to solution, and many ways have been proposed for deriving the true flow curve from experimental data.[2,3,5–10]

The method proposed here is an approximation which is very simple to apply, fast and generally extremely accurate. The equations used to calculate shear rates and shear stresses are the same as those used for Newtonian flow, with the exception that they are applied at chosen intermediate radii rather than at the viscometer walls. Although the flow curve derived by this method is, except in special cases, only an approximation to the true curve, the accuracy may often be well within the limits of precision of the experimental data. The method is particularly suitable for interpreting rotational-viscometer data, first because of the inherently high accuracy in this case, and second because the graphical differentiation methods are lengthy and relatively sensitive to errors in determination of slopes. In the case of the capillary viscometer, use of the proposed method is relatively less advantageous and probably is warranted only in one or more of the following cases: (a) when deviation from Newtonian flow is small, (*b*) when a logarithmic plot of the flow curve is linear, or (c) when many similar curves must be determined.

The method has been worked out in detail only for pseudo-plastic flow; however, the principle could be applied equally well to any other type of non-Newtonian flow.

PRINCIPLE OF THE METHOD

At some particular intermediate radius in either a capillary viscometer or a rotational viscometer the rate of shear calculated from the apparent Newtonian viscosity is the same as the true rate of shear. Line *OABCD* in Fig. 1 is a typical pseudo-plastic flow curve. Consider the flow of such a liquid through a tube under such applied pressure gradient that the shear stress at the wall of the tube is τ_R. The shear-rate profile in

* This article is based upon part of a dissertation to be submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Princeton University.

† Research Fellow, Textile Research Institute. Present address, Chemical Division, General Electric Company, Pittsfield, Massachusetts.

‡ Consultant, Textile Research Institute, and Professor of Chemical Engineering, Princeton University.

[1] B. Rabinowitsch, Z. physik. Chem. **A145**, 1 (1929).
[2] I. M. Krieger and S. H. Maron, J. Appl. Phys. **25**, 72 (1954).
[3] M. Mooney, J. Rheol. **2**, 210 (1931).
[4] O. H. Clark and M. L. Deutsch, J. Appl. Phys. **21**, 713 (1950).
[5] W. Philippoff, *Viskosität der Kolloide* (Theodor Steinkopff, Dresden, 1942; and Edwards Brothers, Inc., Ann Arbor, Michigan, 1944).
[6] J. Pawlowski, Kolloid-Z. **130**, 129 (1953).
[7] F. Schultz-Grunow and H. Weymann, Kolloid-Z. **131**, 61 (1953).
[8] W. Tillmann, Kolloid-Z. **131**, 66 (1953).
[9] C. G. Lindquist and W. C. Sierichs, J. Colloid Sci. **6**, 33 (1951).
[10] R. Wratten, *Proceedings of the Second International Congress on Rheology*, edited by V. G. W. Harrison (Academic Press, Inc., New York, 1954), p. 181.

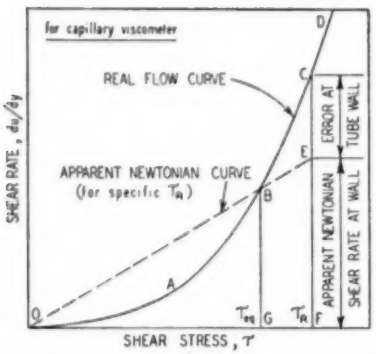


FIG. 1. Plot of typical pseudo-plastic flow curve demonstrating principle of equivalent-radius method for flow in a tube. *OABCD*—flow curve, *OBE*—apparent Newtonian flow curve for particular wall stress τ_R, τ_R—shear stress at wall of tube, $_{eq}$—shear stress at radius r_{eq} where real and apparent curves cross, *FC*—real shear rate at tube wall, i.e., at stress τ_R, *FE*—apparent shear rate at wall calculated from equivalent apparent Newtonain viscosity, *CE*—error resulting from use of apparent Newtonian viscosity in calculating shear rate at wall, *GB*—shear rate at r_{eq} where real and apparent curves cross and real and apparent shear rates are equal.

the tube will be similar to that part of the flow curve from *O* to *C*. If, by the use of Poiseuille's equation, an apparent viscosity is calculated from the measured flow rate, pressure gradient, and tube dimensions, the equivalent apparent Newtonian flow curve, line *OBE*, can be plotted since its slope is the reciprocal of the apparent viscosity. If the volumetric flow rates derived from both curves are to be equal, the two curves, real and apparent, must cross at some stress value τ_{eq}, lower than τ_R. The shear stress τ_{eq} obtains at some intermediate radius r_{eq}. Shear rates for points on the apparent flow curve can be calculated by dividing shear stresses for particular radii by the apparent Newtonian viscosity. At radius *R*, a point so calculated would lie below the true curve by the distance labelled "Error" in Fig. 1, but at radius r_{eq} a point so calculated would lie on the true curve. Thus when r_{eq} is known, Poiseuille's equation, or the corresponding rate and stress equations for Newtonian flow, can be used to calculate true flow curves of non-Newtonian liquids.

The more rigorous differentiation methods yield correct shear rates at the wall. In effect, the errors incurred by use of the Newtonain equations are computed and used to correct the apparent Newtonian shear rates at the wall radius. The equivalent-radius method yields an approximation to the correct shear rate at an intermediate radius. The principal purpose of this paper is to show how this intermediate radius can be estimated and used. The accuracy of the approximation will also be considered. It will be shown that, although the specific value of r_{eq} depends upon the form of the flow curve and may vary with the flow rate, the variation is small, and a single properly chosen value will usually suffice.

In the case of the rotational viscometer the treatment is similar and Fig. 2 is self-explanatory, although there is one significant difference between the two cases. For the capillary viscometer the equivalent apparent shear rate is a complex average of the entire shear-rate range from zero stress to τ_R, the shear stress at the wall. For the rotational viscometer the range average is from τ_c to τ_b, shear stress at the cup wall to shear stress at the bob wall. Because of the narrower range, and also because the functionality is different, closer agreement between real and apparent curves is obtained at either wall, and very much closer agreement is obtained at a properly chosen intermediate radius. The problem is to select the proper intermediate radius.

THEORY

List of symbols:

ζ = viscosity counterpart for non-Newtonian flow
τ = shear stress
Ω = angular speed of cup relative to bob
b = bob radius
c = cup radius
du/dy = shear rate
K_1, K_2, K_3, K_4 = constants defined by Eqs. (3A) and (7A) or by (8A) and (12A)
L = length of capillary (capillary viscometer)
L = length of bob (rotational viscometer)
$[NN/N]$ = ratio of du/dy calculated by power-flow law to du/dy calculated by Newtonian flow law
P = pressure drop through capillary
Q = volumetric flow rate through capillary
R = radius of capillary
r = intermediate radius in either type viscometer
r_{eq} = radius at which du/dy calculated by Newtonian equation is identical with true du/dy

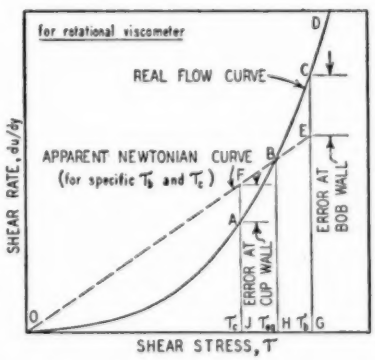


FIG. 2. Plot of typical pseudo-plastic flow curve demonstrating principle of equivalent-radius method as applied to a rotational viscometer. *OABCD*—flow curve, *OFBE*—apparent Newtonian flow curve for particular wall stresses τ_b and τ_c, τ_b—shear stress at bob wall, τ_c—shear stress at cup wall, τ_{eq}—shear stress at r_{eq}, *JA*—real shear rate at cup wall, *JF*—apparent shear rate at cup wall calculated from equivalent apparent Newtonian viscosity, *AF*—error resulting from use of apparent Newtonian viscosity, *HB*—shear rate at r_{eq} where real and apparent curves cross and real and apparent shear rates are equal.

s = radius ratio c/b
T = torque
$\chi = r/R$, used to specify r
Z = stress exponent in power function, Eq. (2), or slope of log-log flow curve.

The flow curve defines the relationship between shear rate and shear stress during steady-state flow, i.e.,

$$du/dy = f(\tau). \tag{1}$$

For any specific form of viscometer, integration of this equation, using the proper boundary conditions, yields the relationship between the measured variables, for instance between pressure gradient and flow rate for a capillary viscometer.

Flow curves of high-polymer solutions are frequently fitted by the empirical equation[3,5,11–15]

$$du/dy = (1/\zeta)\tau^z. \tag{2}$$

The method developed in this paper is based upon the use of this simple power function.

CAPILLARY VISCOMETER

For a Newtonian liquid the shear rate at radius $r = \chi R$ is§

$$\left(\frac{du}{dy}\right)_N = \frac{4Q\chi}{\pi R^3}, \tag{3}$$

and for a non-Newtonian liquid, of the type specified by Eq. (2), the shear rate is

$$\left(\frac{du}{dy}\right)_{NN} = \frac{(Z+3)Q\chi^Z}{\pi R^3}. \tag{4}$$

For a particular flow rate, the radius r_{eq} at which shear rates calculated by the two methods are identical can be found by equating Eqs. (3) and (4) and solving for χ. The *equivalent radius*, thus obtained, is

$$r_{eq} = \left(\frac{4}{Z+3}\right)^{1/(Z-1)} R. \tag{5}$$

At any radius the ratio of the shear rate calculated by the power-function method to that calculated by the Newtonian method is

$$[NN/N]_{CAP} = \frac{(du/dy)_{NN}}{(du/dy)_N} = \frac{(Z+3)}{4}\chi^{Z-1}. \tag{6}$$

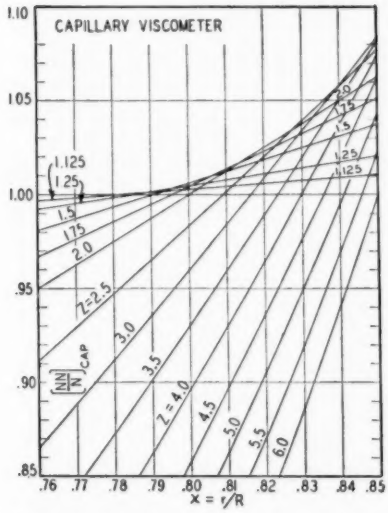


FIG. 3. Capillary viscometer. Ratio of power-function shear rate to equivalent Newtonian shear rate.

Equation (6) defines the basic function $[NN/N]_{CAP}$. Figure 3 is a plot of this function against χ at various values of Z.

The shear stress at radius r is independent of the form of the flow curve and is

$$\tau_r = \frac{Pr}{2L}. \tag{7}$$

ROTATIONAL VISCOMETER

For a Newtonian liquid the shear rate at radius r, where $b \leqq r \leqq c$, is

$$\left(\frac{du}{dy}\right)_N = \frac{2c^2\Omega}{r^2(s^2-1)}, \tag{8}$$

and for a non-Newtonian liquid, of the type specified by Eq. (2), the shear rate is

$$\left(\frac{du}{dy}\right)_{NN} = \frac{2Z\Omega}{(s^{2Z}-1)}\left(\frac{c}{r}\right)^{2Z}. \tag{9}$$

Proceeding as before, and dividing the equivalent radius thus obtained by $(bc)^{\frac{1}{2}}$, one obtains

$$\frac{r_{eq}}{(bc)^{\frac{1}{2}}} = s^{\frac{1}{2}}\left(\frac{Z(s^2-1)}{s^{2Z}-1}\right)^{1/(2Z-2)}. \tag{10}$$

Figure 4 is a plot of Eq. (10).

At any intermediate radius, r, the ratio of the shear rate calculated by the power-function method to that calculated by the Newtonian method is

$$[NN/N]_{RV} = \frac{(du/dy)_{NN}}{(du/dy)_N} = \frac{Z(s^2-1)}{(s^{2Z}-1)}\left(\frac{c}{r}\right)^{2Z-2}. \tag{11}$$

[11] I. M. Krieger and S. H. Maron, J. Colloid Sci. **6**, 528 (1951).
[12] M. Reiner, *Deformation and Flow* (H. K. Lewis, London, 1949).
[13] Farrow, Lowe, and Neale, J. Textile Inst. **19**, T18 (1928).
[14] J. G. Krieble and J. C. Whitwell, Textile Research J. **19**, 253 (1949).
[15] G. Guy Barr, *A Monograph of Viscometry* (Oxford University Press, London, 1931).
§ The development of the basic equations can be found in any text on viscometry (see references 5, 12, and 15).

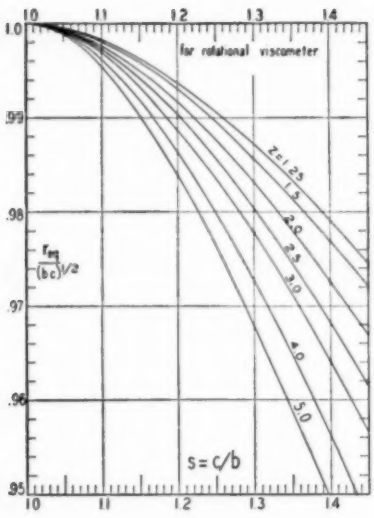


FIG. 4. Rotational viscometer. Graph for determination of equivalent radius.

Equation (11) contains one more variable than did Eq. (6) so that a general plot applicable to all rotational viscometers, similar in use to Fig. 3, cannot be constructed. However, if a specific viscometer is considered, where s is fixed, construction of a single plot is feasible. One based on Eq. (11), for the authors' viscometer was constructed and found useful but is not presented here since it is specific to the one instrument.

If r is expressed in terms of s at constant Z, another plot of more general use can be made. Figure 5, applicable to any viscometer, has been constructed by setting $Z=2$, calculating the corresponding value of r_{eq} in terms of s and c, and then substituting this particular r_{eq} into Eq. (11). The value of $Z=2$ was chosen for general utility as will be noted later.

In all cases the shear stress at radius r is given by

$$\tau_r=\frac{T}{2\pi Lr^2}. \tag{12}$$

The basic equations for preparation of the graphs and calculation of flow curves have been presented in this section. Methods of computing flow curves, and estimating their accuracy, follow.

For Newtonian fluids shear rate and shear stress at any radius, $r=\chi R$ can be computed from Eqs. (3) and (7) for capillary viscometers and Eqs. (8) and (12) for rotational viscometers. With fixed dimensional constants, these equations may be simplified to

$$(du/dy)_r=K_1Q, \tag{3A}$$

$$\tau_r=K_2P, \tag{7A}$$

$$(du/dy)_r=K_3\Omega, \tag{8A}$$

and

$$\tau_r=K_4T. \tag{12A}$$

Substitution of viscometric data into these equations provides the coordinates of points on the true flow curves.

For non-Newtonian fluids absolute flow curves may be calculated from the same equation if the equivalent radius is substituted for the true radius. The steps in the calculation are as follows:

1. A value of Z is assumed and r_{eq} is obtained from Fig. 3 or Fig. 4, or from Eqs. (5) or (10).

2. Dimensional constants K_1 and K_2, or K_3 and K_4, are calculated and from these are obtained at least two corresponding values of shear rate and shear stress on the flow curve.

3. The logarithms of the calculated shear rate and shear stress points are plotted and used to obtain a check on the originally assumed value of Z.

4. If the measured Z varies appreciably from that assumed, estimate the error caused by the original assumption using Figs. 3 or 5 and recalculate r_{eq} if necessary. Since viscometric data are often not extremely precise, it is unlikely that this recalculation will be required.

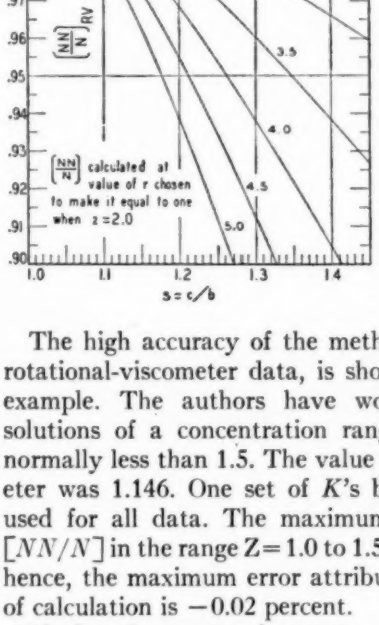


FIG. 5. Rotational viscometer. Ratio of power-function shear rate to equivalent Newtonian shear rate.

The high accuracy of the method, when applied to rotational-viscometer data, is shown by the following example. The authors have worked with cellulose solutions of a concentration range such that Z was normally less than 1.5. The value of s for their viscometer was 1.146. One set of K's based on $Z=1.5$ was used for all data. The maximum value of the ratio $[NN/N]$ in the range Z=1.0 to 1.5 is 1.0002 at $Z\cong 1.25$; hence, the maximum error attributable to the method of calculation is -0.02 percent.

If, for the same viscometer with $s=1.146$, Z is assumed to be 2.0 in calculation of the K's then $[NN/N]=1.0007$ for a true Z of 1.5. The error thus incurred is -0.07 percent. If, in this same case, shear rates were calculated at the bob or cup walls the corresponding errors would be $+6.7$ percent and -7.2 percent, respectively.‖

‖ E. K. Fischer and C. H. Lindsley [J. Colloid Sci. 3, 111 (1948)] realized that an effect of this type existed and reduced the extent of the dependence of their curves upon their particular instruments by use of a calculational procedure that amounted to application of the Newtonian equations at $r=(bc)^{\frac{1}{2}}$.

TEST OF THE METHOD

To test the method, flow curves were calculated from the data of Krieger and Maron[2] and compared with those obtained by them using their more rigorous differentiation method. Krieger and Maron, working with a 62.2 percent solids GR-S Latex, used one capillary and two rotational viscometers to obtain their data. Their data, viscometer dimensions, and methods of calculation are reported in detail in their paper. The rotational-viscometer cup-to-bob ratios, s, were 1.064 and 1.113 for viscometers A and B, respectively.

To derive flow curves from the same data by the equivalent-radius method, the authors used $Z=2$, calculated the corresponding K's from Eqs. (3A) and (7A) for the capillary viscometer, and (8A) and (12A) for the rotational viscometers, then used these equations to compute the complete flow curves from Krieger and Maron's original data. Since Krieger and Maron plotted their results in the form $g(F)/F$ *vs* F, where $g(F)$ is shear rate at shear stress F, the ratio $(du/dy)/\tau$ was calculated and plotted *vs* τ in Fig. 6. Krieger and Maron's results have been plotted on the same figure for comparison. Logarithmic coordinates rather than linear coordinates, as in Fig. 3 of Krieger and Maron's paper, have been used to permit better comparison of the relative accuracy over the entire shear-stress range.

On Fig. 6 the slope increases from 0.5 at the top to 2.0 at the bottom. The corresponding slopes, values of Z, on a flow curve are 1.5 and 3.0. A single set of constants was used for each viscometer over the entire range of the data, based on $Z=2$ as just explained. The magnitude of the error incurred due to the calculational method will be a maximum at $Z=3$ and may be estimated from Fig. 6. For Krieger and Maron's viscometer A, $[NN/N]=0.999$ at $Z=3$ and $s=1.064$. For their viscometer B, $[NN/N]=0.996$ at $Z=3$ and $s=1.113$. The shear rates computed by the equivalent-radius method in this example are therefore 0.1 percent high and 0.4 percent high for viscometers A and B, respectively, when $Z=3$. When $Z=2$ they are correct. When Z is less than 2 the error is not greater than about 0.05 percent. The magnitude of these errors is obviously of negligible significance in this type of work.

In Fig. 6, with the exception of the two lowest points of the capillary-viscometer data, the agreement between the curves calculated by the equivalent-radius method and the corresponding curves calculated by Krieger and Maron's differentiation method is substantially better than the agreement between Krieger and Maron's three curves obtained with three different instruments. In the case of the rotational viscometers no difference can be observed between the curves calculated by Krieger and Maron's method and those calculated by the equivalent-radius method.

The fact that the rotational-viscometer curves diverge at low stress values and that the two methods of calculation yield the same results for each instrument indicates an instrument error or a difference in the slurries. That the two different methods of calculation yield identical curves in each case is evidence that the discrepancy is in the experiment rather than in the method of calculation.

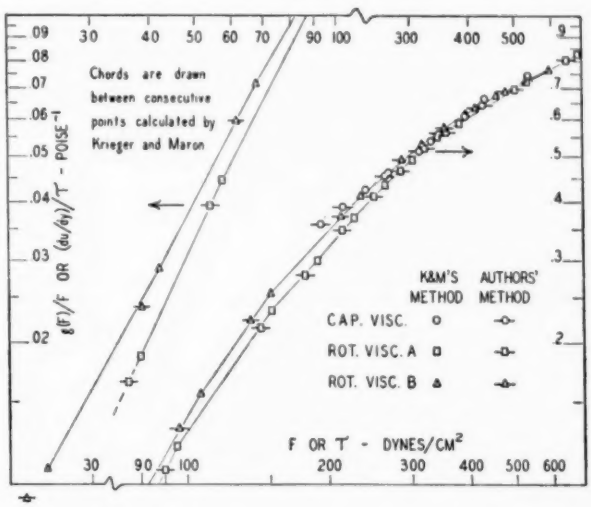

Chords are drawn between consecutive points calculated by Krieger and Maron
g(F)/F OR (du/dy)/τ - POISE
F OR τ - DYNES/CM²
K&M'S METHOD
AUTHORS' METHOD
CAP. VISC.
ROT. VISC. A
ROT. VISC. B

FIG. 6. Plot of point fluidity *vs* shear stress for Krieger and Maron's data (see reference 2) with points calculated by Krieger and Maron's single-bob method and capillary-viscometer method and by the equivalent-radius method.

CONCLUSION

The method proposed in this paper for determining non-Newtonian flow curves from viscometric data reduces the calculation procedure to one similar to that used to calculate shear rates and shear stresses of Newtonian liquids. Graphs have been presented to permit determination of essential instrument constants, after which the procedure consists merely of multiplying the experimentally measured force and rate variables by these constants. In the case of the rotational viscometer, the accuracy obtainable is very high and results can be obtained which are at least as good as those from the theoretically more rigorous but relatively cumbersome graphical differentiation methods. In the case of the capillary viscometer the accuracy is not quite as high, and the graphical differentiation method is relatively straightforward, so use of the equivalent-radius method is relatively less rewarding, but is quicker and easier.

The precision obtainable with this method depends upon the fact that when flow curves of widely different non-Newtonian liquids are plotted on such a scale that the equivalent apparent Newtonian curves of each coincide, then all the flow curves cross the Newtonian curve at almost the same point. A substantial change in the form of the Newtonian curve shifts the common point very little. The graphs presented permit determination of this point, and the method consists

essentially of calculating apparent Newtonian shear rates at this point. At the common point apparent Newtonian shear rates are identical with real share rates.

ACKNOWLEDGMENT

The research upon which this paper is based was made possible by a fellowship granted to K. H. Cram by the Textile Research Institute.

JOURNAL OF APPLIED PHYSICS VOLUME 26, NUMBER 5 MAY, 1955

Resonant-Cavity Field Measurements

S. W. KITCHEN* AND A. D. SCHELBERG†
Radiation Laboratory, Department of Physics, University of California, Berkeley, California
(Received August 21, 1954)

The application of perturbation techniques to the quantitative measurement of both relative electric and relative magnetic fields in resonant cavities is described. The apparatus, procedures, advantages, and limitations are discussed, and the experimental results are compared with the calculable field distributions of a coaxial resonator. The theoretical shunt impedance and Q obtained empirically agree, respectively, within one and four percent of the calculated values.

ALTHOUGH the field distribution in a resonant cavity can be, and has been, calculated for many geometries,[1] frequently the field distribution of a cavity with nonstandard geometry is desired. A well-known technique is measurement with probes. A more recent technique has been the use of small perturbations which cause a frequency shift related in a known way to the local field.[2] A comprehensive examination of these relationships between the fractional frequency shift, the local relative field strength, and the shape of the perturbing body has recently been published by Mullet,[3] who considers dielectric as well as metallic bodies. The application of the technique, however, has apparently been confined to determinations of electric field. This report describes the application of the technique in the 200–400 Mc range to the measurement of both electric and magnetic local fields in resonant cavities with an attainable precision of better than one percent. The apparatus, procedures, advantages, and limitations of the method are discussed, and experimental results are compared with the calculable field distributions of a coaxial resonator.

The general expression relating the frequency shift[4] due to a sphere of permeability μ and dielectric constant ϵ to the local field $(E+H)$ is

$$\frac{\delta\omega}{\omega}=-\frac{3}{4}\frac{\delta v}{W}\left[\frac{\epsilon-1}{\epsilon+2}\epsilon_0E^2+\frac{\mu-1}{\mu+2}\mu_0H^2\right], \tag{1}$$

where δv is the volume of the sphere and W is the stored energy. For a dielectric sphere, this equation may be simplified to:

$$\left(\frac{\delta\omega}{\omega}\right)_D=-\frac{3}{4}\delta v_D\left(\frac{\epsilon-1}{\epsilon+2}\right)\epsilon_0\frac{E^2}{W}. \tag{2}$$

Consequently, a dielectric sphere may be used to determine the relative magnitude of electric field even in the presence of a magnetic field. In the known absence of magnetic field, a conducting sphere will produce a fractional frequency shift:

$$\left(\frac{\delta\omega}{\omega}\right)_M=-\frac{3}{4}\frac{\delta v_M}{W}\epsilon_0E^2. \tag{3}$$

The ratio of Eq. (3) to Eq. (2) at the same point in the cavity is

$$\frac{(\delta\omega/\omega)_M}{(\delta\omega/\omega)_D}=\frac{\epsilon+2}{\epsilon-1}\frac{\delta v_M}{\delta v_D}=\xi \tag{4}$$

for a given pair of metal and dielectric spheres. The value of ξ can be determined empirically and used to determine the magnetic-field distribution. In the presence of both magnetic and electric fields

$$\left(\frac{\delta\omega}{\omega}\right)_M=-\tfrac{3}{4}\frac{\delta v_M}{W}[\epsilon_0E^2-\tfrac{1}{2}\mu_0H^2]$$

* Now at Knolls Atomic Power Laboratory, Schenectady, New York.

† Now at Los Alamos Scientific Laboratory, Los Alamos, New Mexico.

[1] R. D. Hill, Sci. Inst. **22**, 221 (1945); "Microwave transmission design data," Sperry Gyroscope Company, 1944; S. A. Schelkunoff, *Electromagnetic Waves* (D. Van Nostrand Company, Inc., New York, 1943); W. C. Hahn, J. Appl. Phys. **12**, 62–64 (1941); S. Ramo and J. R. Whinnery, *Fields and Waves in Modern Radio* (John Wiley and Sons, Inc., New York, 1944).

[2] W. W. Hansen and R. F. Post, J. Appl. Phys. **19**, 1095 (1948); L. C. Maier, Jr., and J. C. Slater, J. Appl. Phys. **23**, 68 (1952); J. Müller, Hochfreq. Tech. and Elektroakustik **54**, 157 (1937); H. A. Bethe and J. Schwinger, MIT Radiation Laboratory Report D1-117, March 4, 1943; H. B. G. Casimir, Philips Research Repts. **6**, 162 (1951); A. Gilardini, J. Appl. Phys. **25**, 1064 (1954).

[3] L. B. Mullet, "Perturbation of a resonator," A. E. R. E. G/R 853 (1952).

[4] Note that $\delta\omega/\omega$ measures only the ratio of the local fields relative to the average fields in the cavity, and is therefore independent of the power consumption. Owing to the difficulties of measuring power precisely, this factor is an important advantage of the perturbation technique.

and

$$\frac{H^2}{W}=\frac{8}{3\mu_0\delta v_M}\left[\left(\frac{\delta\omega}{\omega}\right)_M-\xi\left(\frac{\delta\omega}{\omega}\right)_D\right]. \tag{5}$$

With the aid of these relations all the theoretical constants of a cavity, such as shunt resistance (Z_s), transit time factor (T), or (Q) can be determined with an accuracy dependent principally on the measurement of ($\delta\omega/\omega$). For example, Z_s, Q,[5] and T may be expressed as:

$$Z_s=\frac{\frac{\mu_0}{\epsilon_0}\sigma\Delta\left[\int(\delta\gamma)^{\frac{1}{2}}ds\right]^2}{2\int[\xi(\delta\gamma)_D-(\delta\gamma)_M]dA}\text{ ohms}$$

$$\frac{1}{Q}=\frac{2\Delta}{3\delta v_M}\int\left[\left(\frac{\delta\omega}{\omega}\right)_M-\xi\left(\frac{\delta\omega}{\omega}\right)_D\right]dA \tag{6}$$

$$T=\frac{\int(\delta\gamma)^{\frac{1}{2}}\cos(\omega t+\phi)ds}{\int(\delta\gamma)^{\frac{1}{2}}ds},$$

where $\int Fds$ is a line integral, $\int FdA$ is a surface integral, Δ is the skin depth, γ is the frequency, and all units are rationalized MKS units. The difference terms in Z_s and Q limit the accuracy of these quantities as obtained from the perturbation technique, even though the apparatus described below permits measurement of $\delta\omega/\omega$ to 10^{-8}.

One difficulty encountered in measuring $\delta\omega/\omega$ for substitution in the above formulas lies in the fact that the perturbation formula (1) does not apply to spheres that are close to a conducting surface, for then image effects come into play. In practice, $\delta\omega/\omega$ observed when the sphere is less than a diameter away from a metal surface is discarded, although it is possible to calculate a proximity correction by the following expressions for the measured fields in terms of the actual fields. Thus

$$E=E_a\left[1+\frac{1}{4}\frac{\epsilon-1}{\epsilon+2}\left(\frac{r}{d}\right)^3\right]\quad(\text{for metal }\epsilon\to\infty)$$

$$H=H_a\left[1+\frac{1}{16}\left(\frac{r}{d}\right)^3\right], \tag{7}$$

where r is the radius and d is the distance from the center of the sphere to the surface. These expressions were derived by considering the dipole image effects.

Where reasonably accurate extrapolations to the surface are not possible, another technique may be used. It can be shown that a hemisphere on a plane surface will produce half the frequency shift produced by a sphere of the same radius in the same field.[6] The difficulty encountered in sliding a hemisphere along a surface is that the surface is rarely plane. If the hemispherical diameter is small compared to the radius of curvature of the surface, however, satisfactory data can be obtained by calibrating the ratio of spherical to hemispherical shift in a cavity of standard-field distribution, even though the ratio for metal spheres will have one value for the shift due to the magnetic field and another for the shift due to the electric field.

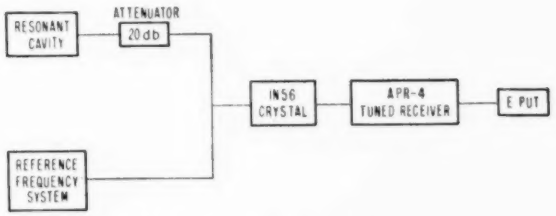


FIG. 1. Frequency measuring system.

A block diagram of the frequency measuring system is shown in Fig. 1. The frequency of the reference system[7] is adjusted so that it is lower than the cavity frequency by about 1500 cps. The cavity itself is driven by a grid-dip oscillator. Choice of a grid-dip type of oscillator was dictated by the fact that the grid-dip frequency is controlled over the range of $\delta\gamma$ by the cavity. The perturbing sphere (or hemisphere) is introduced into the cavity on a thread, $\delta\gamma$ being observed on an EPUT.[8]

The accuracy of this method of measuring relative field distribution is determined by several factors:

(1) the stability of the reference frequency systems,
(2) mechanical stability of the cavity,
(3) the magnitude of perturbances other than the desired one during the measurement,
(4) errors in positioning the sphere,
(5) "following" of the cavity resonant frequency by the cavity oscillator.

The requirements were much more severe for the measurements of magnetic field than those of electric field.

The time required to make one measurement of $\delta\gamma$ is about ten seconds. The reference frequency system is stable to less than ten cycles per second over a ten-second period at worst, and frequently drifts no more than ten cycles over a period of several minutes. Possible error due to this drift is negligible compared to the other factors.

Another source of error is a change in the resonant frequency of the cavity due to thermal expansion; drifts as high as 100 cps per second have been observed.

[5] Originally derived by M. L. Good of this Laboratory.

[6] J. C. Slater, Revs. Modern Phys. **20**, 473 (1948).

[7] R. G. Smits and A. J. Schwemin, "A stable microwave reference system," University of California Radiation Laboratory Report No. UCRL-1843, June, 1952.

[8] Events Per Unit Time Meter, Berkeley Scientific Corporation, Richmond, California.

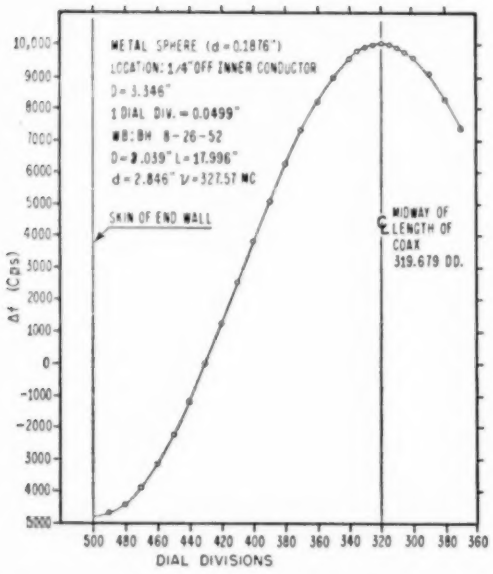

METAL SPHERE (d = 0.1876")
LOCATION: 1/4" OFF INNER CONDUCTOR
D = 3.346"
1 DIAL DIV. = 0.0499"
WB:BH 8-26-52
D = 8.039" L = 17.996"
d = 2.846" ν = 327.57 MC
SKIN OF END WALL
℄ MIDWAY OF LENGTH OF COAX 319.679 DD.
Δf (Cps)
DIAL DIVISIONS

FIG. 2. Frequency shifts due to mixed electric and magnetic fields.

For precise measurements, therefore, it was found necessary to control the temperature of the room where the cavity was under observation to $\pm 0.5°F$ as recorded with a thermocouple thermometer. Although the magnitude of the effect was never determined, the drift was further reduced by running the grid-dip oscillator with dc on the filaments and leaving all oscillators on continuously.

Mechanical vibration of a cavity during a measurement was another source of error. Stiffening ribs around the precision cavity's periphery were found to be advisable even through the wall was $\frac{1}{4}$ inch thick. In addition, the test cavities were mounted on Lord shock mounts as further protection against vibration. The sensitivity of the frequency measuring system, however, was such as to respond even to thumb pressure against the cavity. As a result, the guide pulleys had to be mounted in such a way as to prevent stress on the cavity when positioning the sphere.

The thread supporting the sphere had two requirements to meets: (1) that its nonuniformity be sufficiently low that frequency shifts due to the thread alone could be kept much less than those due to the sphere, and (2) that its modulus of elasticity be high. Of all available types of thread—cotton, silk, linen, glass, and nylon; twisted, braided, and monofilament; some impregnated and some plastic coated—by far the most uniform were nylon monofilament and an impregnated silk suture with the trade name of Dermal. The least uniform was linen. So far as stretch is concerned, however, glass and linen were the best and nylon was by far the worst. The principal drawback to the glass thread was its inability to withstand stress around sharp corners. Dermal, when boiled and prestretched, was entirely satisfactory in all respects. For less precise measurements, braided silk would be adequate. Satisfactory measurements of magnetic field in the presence of electric field require, however, that the metal and dielectric spheres be at the same point in the cavity, particularly when the field strengths are a rapid function of position.

Another source of possible difficulty lies in the method of exciting the cavity. Even though the cavity has a very high Q as compared to the grid-dip oscillator, it is still a tuned circuit coupled to another tuned circuit. Consequently, the observed resonant frequency is not exactly the resonant frequency of the cavity. Furthermore, as the cavity resonant frequency changes owing to perturbation, the change in the excitation frequency observed in the cavity is not equal to the change in the cavity resonant frequency, but depends on the setting of the grid dip. It was found that for a setting nearly equal to the maximum needle dip, the ratio of $\delta\omega$ observed to $\delta\omega$ actual would remain constant at about 0.95 for shifts up to 0.05 mc. For values of Z_s and T, the exact value of the ratio is unimportant as long as one knows it remains constant over the range of $\delta\gamma$ for a given setting of the grid dip. Calculation of Q, however, requires a measurement of the ratio using a standard geometry.

The determination of Q requires, in addition to the absolute frequency shift, a knowledge of the shape and diameter of the sphere. Bronze ball bearings make excellent metal spheres, but no satisfactory dielectric spheres could be procured until the technique of making precise fused quartz spheres was developed.[9]

To illustrate the quality of the technique, Figs. 2–4 show the measurements made on the coaxial cavity of Fig. 5. Note that the agreement between the observed and calculated distributions is excellent on

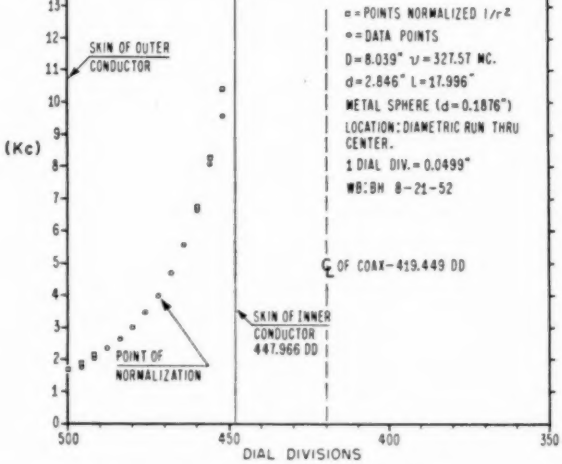

α = POINTS NORMALIZED 1/r²
∘ = DATA POINTS
D = 8.039" ν = 327.57 MC.
d = 2.846" L = 17.996"
METAL SPHERE (d = 0.1876")
LOCATION: DIAMETRIC RUN THRU CENTER.
1 DIAL DIV. = 0.0499"
WB:BH 8-21-52
SKIN OF OUTER CONDUCTOR
℄ OF COAX - 419.449 DD
SKIN OF INNER CONDUCTOR 447.966 DD
POINT OF NORMALIZATION
(Kc)
DIAL DIVISIONS

FIG. 3. Radial variation of electric field squared in arbitrary units.

[9] A. J. Schwemin, "Making precise nonmetallic spheres," University of California Radiation Laboratory Report No. UCRL-1560, November, 1951; W. L. Bond, Rev. Sci. Instr. **22**, 344 (1951); P. Senio and C. W. Tucker, Jr., Rev. Sci. Instr. **24**, 549 (1953).

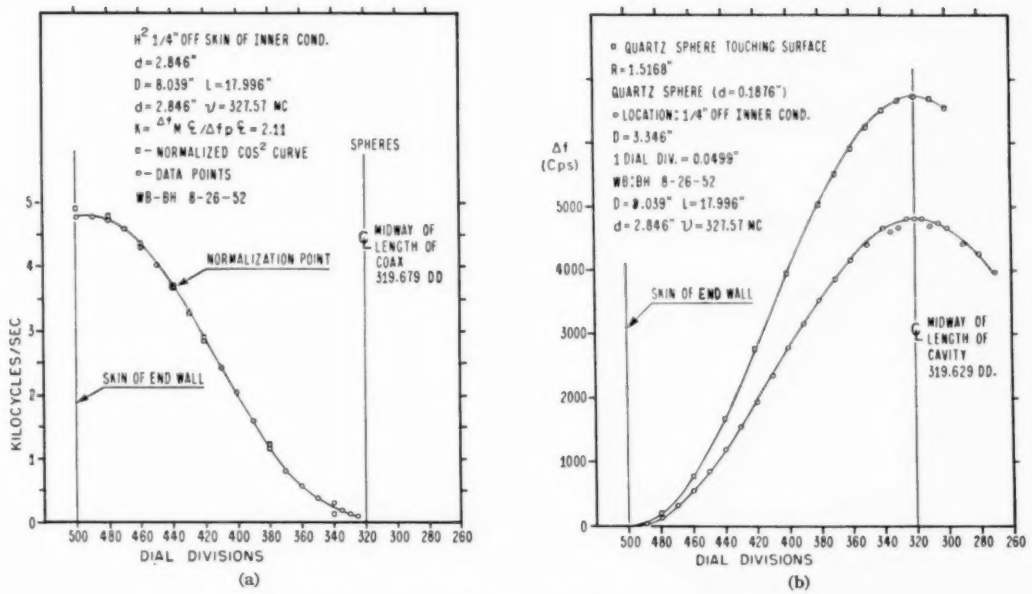


FIG. 4. (a) Axial magnetic field squared. (b) Axial electric field squared.

the whole. The only major deviations occur near surfaces or in regions of very low field. Equation (7) indicates that the imaged dipole acts to increase both the electric field and the magnetic field at the sphere; the observed frequency shifts due to these fields behave in a like manner. This effect was verified by observations taken along the inner conductor with spheres touching the conductor and $\frac{1}{4}$ inch off [see Fig. 4(b)]. The pronounced dip at the upper end of the curve of Fig. 3 seems to be at odds with predicted behavior. These data, however, were taken perpendicular to the surface. Consequently, in this region, the sphere was approaching a $\frac{3}{8}$-inch hole in the conducting surface.

When the aberrations caused by surface-proximity effects were removed by logical extrapolation of the curves as indicated in the figures, the $\delta\omega$ for the spheres was used to calculate the theoretical shunt impedance and theoretical Q of the coaxial cavity with the assumption that the surfaces were of copper whose conductivity was 5.8×10^7 mhos/meter. If the end-wall losses were not included, the Z_s and Q calculated from the data were respectively 6.07×10^5 ohms and 16 000. As calculated from Maxwell's equations, they were 6.02×10^5 ohms and 14 600. The 10 percent discrepancy between the measured and calculated Q values arises principally from the fact mentioned earlier that the $\delta\omega/\omega$ observed is always less than the theoretical value. The measured Q must be corrected by the ratio of the observed $\delta\omega/\omega$ to the theoretical $\delta\omega/\omega$ (equal to 0.95). The Q calculated from the data then becomes 15 200, four percent higher than the predicted value. Even though Q obtained by this method is the least accurate of the quantities obtained, it is still a far better figure for cavities of nonstandard geometry than can be obtained by educated guesses.

In summary, one may say that the perturbation technique may be applied successfully to the quantitative measurement of both magnetic and electric fields within the volume of a resonant cavity. Unlike the probe technique, it does not distort the fields other than locally, and it does not depend on the power level. On the other hand, the accuracy is impaired by small mechanical and thermal changes of the resonant system. It is not the best technique for every circumstance, but it does permit quantitative measurements that were not previously possible.

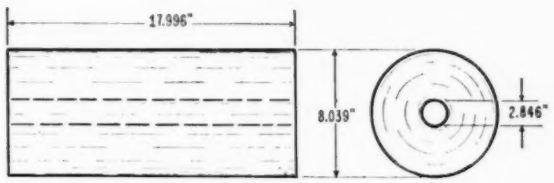


FIG. 5. Coaxial cavity interior dimensions.

It is desired at this time to acknowledge the contributions of Dr. Andrew Longacre, who initiated the program, Dr. W. K. H. Panofsky, who suggested the use of the method for magnetic field measurements, and of A. J. Schwemin and R. G. Smits, whose instrumentation improvements materially added to the speed and accuracy of the measurements.

Noise at the Potential Minimum in the High-Frequency Diode

D. A. WATKINS
Stanford University, Stanford, California
(Received July 29, 1954)

Noise quantities at the potential minimum in the high-frequency diode are analyzed by means of a relatively simple model which assumes that the transit angle from the cathode to the potential minimum is small and that the transit angle from the potential minimum to the anode is large. The mean-square convection current is found to be about half full shot noise for a typical case and the mean-square velocity is found to be the same as in the temperature-limited case. Application of this result to an analysis of the minimum obtainable noise figure of klystrons and traveling-wave tubes using velocity-jump noise reduction shows that their noise figure can be reduced without theoretical limit by increasing the current density at the cathode.

INTRODUCTION

THEORETICAL predictions of the minimum obtainable noise figure for velocity-jump low-noise amplifiers (traveling-wave tubes or klystrons)[1–5] are critically dependent upon the assumptions which are made regarding the noise quantities at the potential minimum in the electron gun. If it is assumed that the cathode is temperature-limited, there is no potential minimum and hence no space-charge smoothing. For this case the noise convection current at the entrance to the diode corresponds to full shot noise and the noise velocity is that calculated by Rack.[6] Since these two fluctuation quantities (convection current and velocity) are uncorrelated, it can be shown that the minimum obtainable noise figure for velocity-jump tubes with temperature-limited oxide cathodes is about 6 db.[7]

Recently Bloom and Peter,[8] Pierce and Danielson,[9] and Robinson[10] have extended this idea by means of linear-transducer theory to include arbitrary dc velocity changes along the electron beam between the cathode and the circuit of a traveling-wave tube and have included the effects of space-charge and loss in the circuit region. They show that loss and space-charge will increase this minimum perhaps 1 db in a practical tube. If the electron gun is not temperature-limited, the noise quantities at the entrance will be different and as a result, a different minimum noise figure for beam-type tubes can be expected.

This paper shows by means of a relatively simple calculation how the noise convection current at the potential minimum differs from full shot noise and how it depends on cathode temperature, dc current density, and cathode-to-potential-minimum distance. An argument is presented which shows that the noise velocity at the potential minimum is unchanged from that associated with full shot noise.[6]

ASSUMPTIONS

In most problems concerning the behavior of high-frequency tubes, and particularly in problems concerning noise, it is necessary to make certain assumptions some of which are difficult to justify. The assumptions made here are as follows:

(1) The three-dimensional aspect of the problem is ignored. The diode is considered to have only one pertinent dimension, that of distance measured in the direction of flow of the electrons (see Fig. 1).

(2) The diode, insofar as its noise behavior is concerned, is considered to be a linear device. This means that the fluctuation quantities are considered to be small compared with the corresponding average or dc values.

(3) The transit angle from the cathode to the potential minimum is assumed to be small—less than π or so.

(4) The transit angle from the potential minimum to the anode is assumed to be large—greater than 4π. This condition is always met in practical tubes.

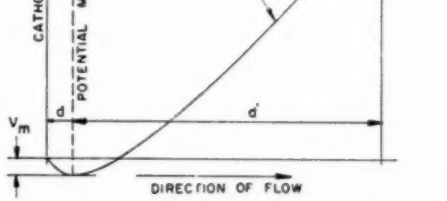


FIG. 1. Schematic representation of the space-charge limited diode showing the form of the dc potential variation and some of the sumbols used.

[1] J. R. Pierce, *Traveling-Wave Tubes* (D. Van Nostrand Company, Inc., New York, 1950).
[2] D. A. Watkins, Proc. Inst. Radio Engrs. **40**, 65 (1952).
[3] R. W. Peter, RCA Rev. **13**, 344 (1952).
[4] R. W. Peter and J. A. Ruetz, RCA Rev. **14**, 441 (1953).
[5] D. A. Watkins, Proc. Inst. Radio Engrs. **41**, 1741 (1953).
[6] A. J. Rack, Bell System Tech J. **17**, 592 (1938).
[7] D. A. Watkins, "Noise reduction in beam-type amplifiers," Ph. D. dissertation; also Electronics Research Laboratory Tech. Rept. No. 31, Stanford University, Stanford, California (1951).
[8] S. Bloom and R. W. Peter, RCA Rev. **15**, 252 (1954).
[9] J. R. Pierce and W. E. Danielson, J. Appl. Phys. **25**, 1163 (1954).
[10] F. N. H. Robinson, J. Brit. Inst. Radio Engrs. **14**, 79 (1954).

CALCULATION

Our procedure is to divide the diode into two diodes which are considered to be in series. The first diode is the space from the cathode to the potential minimum and the second diode is the space from the potential minimum to the anode. The first diode is considered to be a retarding field diode, which it is. It has for small transit angles an ac admittance per unit area (all units are mks) which is given by

$$Y=\frac{eI_D}{kT_c}+\frac{j\omega\epsilon_0}{d}, \tag{1}$$

where I_D is the net dc current density through the diode, and d is the distance from the cathode to the potential minimum. When short circuited for ac, a total rms noise current density corresponding to full shot noise will flow through the retarding field diode because the ac short circuit prevents the production of compensating currents by the potential minimum.[11] Thus the total short-circuit rms noise current density is

$$I_{sc}=(2eI_D\Delta f)^{\frac{1}{2}}. \tag{2}$$

Knowing the short-circuit current and the internal admittance, the assumption of linearity permits us to apply Thevenin's theorem to obtain the open-circuit rms noise voltage

$$V_{oc}=\frac{I_{sc}}{Y}=\frac{(2eI_D\Delta f)^{\frac{1}{2}}}{(eI_D/kT_c)+(j\omega\epsilon_0/d)}. \tag{3}$$

The retarding field diode is now considered to be in series with the second space. As a consequence of the large transit angle between the potential minimum and the anode, we may consider the whole diode to be open circuited for ac. [Actually, we could take the whole diode to be short circuited but because of the relatively high impedance of the second space this would change our final result by a factor which is approximately $(1+d/d')$.] With the whole diode open circuited, the total rms noise current density I, must be zero. Everywhere within the diode, and in particular at the potential mimimum, the total current is the sum of the convection current and the displacement current. Thus,

$$I=i+j\omega\epsilon_0 E=0 \tag{4}$$

and

$$E=-(V_{oc}/d). \tag{5}$$

Therefore,

$$i=j\omega\epsilon_0\frac{V_{oc}}{d}=\frac{j\omega\epsilon_0}{d}\frac{(2eI_D\Delta f)^{\frac{1}{2}}}{\left(\frac{eI_D}{kT_c}+\frac{j\omega\epsilon_0}{d}\right)}. \tag{6}$$

where i is the rms noise convection current density at the potential minimum.

[11] D. O. North, RCA Rev. **4**, 441 (1940).

Taking the square of the magnitude,

$$\frac{i^2}{2eI_D\Delta f}=R^2=\frac{1}{1+(eI_Dd/\omega\epsilon_0kT_c)^2} \tag{7}$$

where R^2 is the amount by which the mean-square noise convection current density at the potential minimum is reduced from full shot noise. There is a temptation to call this quantity Γ^2, but since this is really a different thing this temptation is put down.

In order to make use of the dc diode solution of of Fry[12] and Langmuir,[13] it seems convenient to express R^2 in terms of two variables defined as follows:

$$\xi_d=\left[2\pi^{\frac{1}{2}}\left(\frac{m}{2kT_c}\right)^{\frac{1}{2}}\left(\frac{eI_D}{m\epsilon_0}\right)^{\frac{1}{2}}\right]d, \tag{8}$$

$$a=\frac{\omega}{2\pi^{\frac{1}{2}}(m/2kT_c)^{\frac{1}{2}}(eI_D/m\epsilon_0)^{\frac{1}{2}}}. \tag{9}$$

The first corresponds to the notation of Fry, and the second is a new variable. In terms of these variables

$$R^2=\frac{1}{1+(\xi_d^2/4\pi a^2)}. \tag{10}$$

According to Langmuir's solution, ξ_d has its maximum value of 2.5539 when the ratio of the emitted current to the current passing the potential minimum approaches infinity. Actually, this value will be approximately correct for a ratio of I_{emitted}/I_D equal to 20 or more. Thus, for complete space charge ($I_{\text{emitted}}/I_D>20$),

$$R^2=\frac{1}{1+0.519/a^2}. \tag{11}$$

EXAMPLE

Evaluation of the constants in Eq. (9) gives the following expression for a in terms of more commonly used units

$$a=1.24(10)^{-5}\frac{(T_c,{}^\circ\text{K})^{\frac{1}{2}}(f,\text{ mc/s})}{(I_D,\text{ amp/cm}^2)^{\frac{1}{2}}}. \tag{12}$$

As an example, for $T_c=1020^\circ$K, a frequency of 3000 mc/s, and a current density of 0.1 amp/cm^2, a has the value 0.665 and R^2 is 0.46 for complete space-charge. Thus, the mean-square noise convection current at the potential minimum is typically about half full shot noise for the complete space-charge condition.

VELOCITY FLUCTUATIONS

In order to understand the nature of the effect of space-charge smoothing on the velocity fluctuations, it is necessary to consider the smoothing mechanism

[12] T. C. Fry, Phys. Rev. **17**, 441 (1921).
[13] I. Langmuir, Phys. Rev. **21**, 419 (1923).

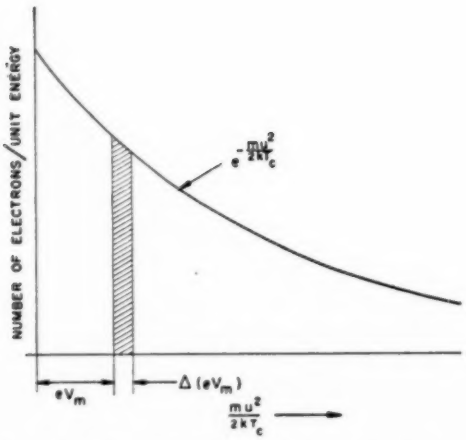


FIG. 2. The Maxwellian velocity distribution at the cathode. The ordinate is proportional to the current carried by electrons having a velocity between $[mu^2/2kT_c$ and $d(mu^2/2kT_c)]$ where u is the velocity. At the potential minimum only electrons whose initial energy exceeds the potential difference eV_m between the cathode and the potential minimum may pass on to the anode. If this potential fluctuates by a small amount $\Delta(eV_m)$ the current passing to the anode will fluctuate but the average velocity of the stream at the potential minimum will not because the average velocity is independent of V_m.

itself. We may think of it in the following physical terms:

Consider that there is a current fluctuation in some velocity class at the cathode of sufficient initial energy to pass the potential minimum. This fluctiation will produce a small fluctuation in the potential at the potential minimum with the result that the Maxwellian distribution function is "chopped off" at a slightly different point as in Fig. 2. Thus, more or less current is released by the potential minimum such that partial compensation for the original fluctuation results. Now here is the heart of the argument. Because of its exponential nature, the Maxwellian distribution has the same *average* velocity no matter where it is chopped off. Thus the fluctuation of the potential minimum adds or subtracts electrons in the lowest velocity class, but at the same time it changes the velocities of all of the other velocity classes so the average velocity of all the electrons passing the potential minimum including the compensating pulse is unchanged by the presence or absence of the compensating pulse.

Rack[6] and others have calculated the mean-square fluctuation of the average velocity of an electron stream under temperature-limited conditions. Since the compensating pulses described in the foregoing are at the average velocity, so to speak, the mean-square fluctuation of the average velocity is unaffected by them. Since the compensation does not affect the mean-square velocity fluctuations, the two entrance quantities at the potential minimum are still completely uncorrelated as they are in the absence of compensation.

Thus, the appropriate entrance quantities for use in calculating noise in high-frequency electron guns starting from the potential minimum are

$$v_a^2=(4-\pi)\frac{e}{m}\frac{KT_c\Delta f}{I_0} \tag{13}$$

and

$$i_a^2=R^2 2eI_D\Delta f \tag{14}$$

where R^2 is given by Eq. (7) or Eq. (10), I_0 is the total beam current, and the cross correlation between the two is zero under the present assumptions.

CATHODE-POTENTIAL MINIMUM TRANSIT ANGLE

Using the Fry-Langmuir dc potential distribution for complete space-charge and a value of a of 0.665, the transit angle for the electron which has the average velocity at the potential minimum is found to be about 1.41 radians or about 80° by numerical integration. This is not small, and slower electrons have longer transit angles. It is not entirely clear what this will do to the result. We might expect that transit angle effects would tend to spoil the compensation, so perhaps we should regard Eq. (11) as a lower limit. Other effects such as a spotty cathode which is partly temperature-limited and partly space-charge limited will also tend to increase R^2 toward unity.

MINIMUM NOISE FIGURE

When the expression for the minimum noise figure of a velocity-jump traveling-wave tube[8-10] is modified to account for the reduction in the noise convection current given here, the result is as follows

$$F_{\min}=1+[\pi(4-\pi)]^{\frac{1}{2}}R\frac{T_c}{T}(4QCf_{\max}f_{\min})^{\frac{1}{2}}. \tag{15}$$

For zero space-charge and zero loss in the circuit region, this becomes

$$F_{\min}=1+(\tfrac{1}{2})[\pi(4-\pi)]^{\frac{1}{2}}R\frac{T_c}{T}. \tag{16}$$

For $R^2=0.46$ and $T_c/T=3.52$, $F_{\min}=2.96$ or 4.7 db.

CONCLUSION

On the assumptions given, we have shown that there is some reduction in shot noise at the potential minimum under the conditions present in guns of velocity-jump low-noise tubes and that under these conditions the velocity fluctuations are the same as in the temperature-limited case. To reduce the minimum noise figure of a traveling-wave tube or klystron it is seen that we should increase the current density at the cathode by keeping the total beam current constant and reducing the cathode diameter. At the same time, the emitted current must be some twenty times the beam current.

The author is indebted to J. R. Pierce of the Bell Telephone Laboratories for suggesting the retarding field diode point of view in 1951. This work was supported jointly by the U. S. Air Force, U. S. Army, and U. S. Navy.

Composition of the Interelectrode Prebreakdown Current in High Vacuum*

HENRY C. BOURNE, JR.†
High Voltage Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts
(Received July 31, 1954)

Thermal rise measurements on vacuum-insulated electrodes subjected to steady gap voltages in the range of 80 to 100 kv show that the interelectrode prebreakdown current is composed chiefly of negatively charged particles. With aluminum electrodes the negative particles outnumber the positive particles (ions) by at least 300 to 1. With steel electrodes, the ratio is greater than 1000 to 1.

INTRODUCTION

IN the study of high-voltage vacuum breakdown much interest has centered on phenomena involving the secondary emission of charged particles from the gap electrodes. In a previous paper[1] the emission of electrons from the cathode caused by positive ion bombardment was investigated. These data, combined with data on positive ion emission caused by electron bombardment, gave some clues as to the role of secondary emission in a theory of high-voltage breakdown in vacuum.[2]

Further insight into the importance of secondary emission in high-vacuum breakdown is obtained by this study of the composition of the interelectrode prebreakdown current. The relative amount of prebreakdown current carried by negatively and positively charged particles is ascertained from temperature measurements at a thermally-sensitive electrode. Although this method reveals only the polarity and energy of the charged particles, it is probable that the negative particles were predominantly electrons and that the positive particles were ions from the anode metal and its residual surface contamination. Because of the short gaps and the low pressures (about 5×10^{-6} mm of Hg) at which these stable prebreakdown discharges were maintained, it is not believed that ionization of the residual gas made a dominating contribution to the positive currents.

In these measurements a Van de Graaff generator is used as the high-voltage source. One electrode in the high vacuum is designed for low heat capacity and is thermally insulated. Welded to its inner surface is an alumel-chromel thermocouple for the temperature measurements. With generator polarity reversal this electrode may be made either the anode or cathode. The other electrode is solidly mounted on an aluminum shaft and purposely given a large heat capacity to reduce radiation between the electrodes. Measurements are made with aluminum electrodes and with steel electrodes.

The net power received by an electrode is proportional to the time-rate of the electrode temperature change. Although thermal power losses by conduction and radiation are minimized, they may be accounted for by adding the thermal power loss to the indicated net power to obtain the power supplied to the electrode by the charged particles. These power losses may be calculated from a measured loss curve, which is taken as the electrode temperature decreases to its ambient value with the power off. The ratio of the thermal power supplied by a particular charged particle to the total electrical gap power is a direct measure of the relative amount of current carried by that particle.

ENERGY TRANSFERRED BY NEGATIVE PARTICLES

In Fig. 1 is shown a temperature-time curve for aluminum electrodes with 100-kv gap voltage and an average power of 0.23 watt. On the same figure is shown the corresponding loss curve.

In order to facilitate the calculations, the temperature curves are approximated very closely by exponential equations. The algebraic difference of the slopes of these two curves give the indicated rate of temperature rise if there were no losses and is equal to 0.91°C/min. With a specific heat of 0.217 calorie per gram for aluminum, the thermal power on the anode is 0.22 watt. Within experimental accuracy this value is indistinguishable from the total electrical gap power of 0.23 watt. This measurement shows that the energy transferred by the negative particles in the interelectrode current is indistinguishable from the total energy transferred. A more precise estimate of the small

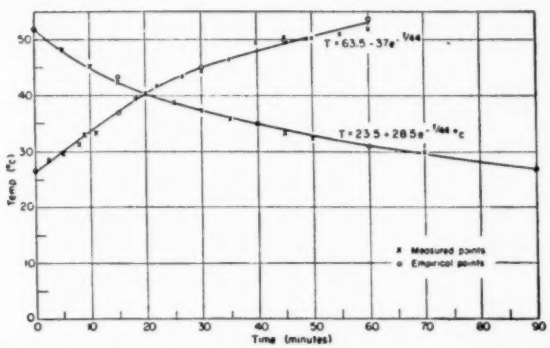


FIG. 1. Rising and falling temperature curves for aluminum anode. Heating by negative particles of interelectrode current. Average total current = 2.3 μa and average total power = 0.23 w. Temperature perception = 0.5°C.

* This work contains part of the results of a thesis submitted by the author in partial fulfillment of the requirements for the degree of Doctor of Science at the Massachusetts Institute of Technology.

† Now faculty member, Division of Electrical Engineering, University of California, Berkeley, California.

[1] Bourne, Cloud, and Trump, J. Appl. Phys. **26**, 596 (1955).

[2] Trump and Van de Graaff, J. Appl. Phys. **18**, 327–332 (1947).

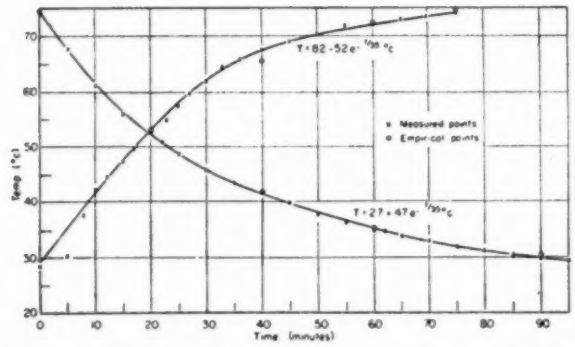


FIG. 2. Rising and falling temperature curves for steel anode. Heating by negative particles of interelectrode current. Average total current = 3.5 μa and average total power = 0.31 w. Temperature perception = 0.5°C.

positive current must be obtained by measurements at the cathode.

The curves of Fig. 2 are taken with steel electrodes. The algebraic difference of the slopes is equal to 1.57°C/min. With a specific heat for the steel electrode of 0.108 calorie per gram, the thermal power on the anode is 0.31 watt, a value identical with the average total gap power. These results, although they do not allow an accurate division of the gap current into its components, do indicate that negative particles carry a large fraction of the gap current.

ENERGY TRANSFERRED BY POSITIVE IONS

With the generator polarity reversed the same measurements are made. The particles impinging on the thermally insulated electrode are now positive ions. The temperature curves for aluminum are shown in Fig. 3. In this case the noticeable oscillations of the temperature are probably due to localized heating and random radiation which are made apparent by the increased sensitivity of the temperature measurements.

Empirical exponential curves are derived for these data also, although they might be equally well represented by straight lines. In either case, the indicated maximum power carried by the positive ions is only 2.8 milliwatts. The positive ions at most can account for only 0.3 percent of the interelectrode current under these conditions.

With steel electrodes, two sets of data are recorded under slightly different conditions. In one case, a $\frac{1}{8}$-inch gap at 60-kv gap voltage and 17-microampere gap current is used. In the other, a $\frac{1}{4}$-inch gap at 150 kv and 7 microamperes is used. In neither case is there any indicated temperature rise over a two-hour observation period. If we assume, within the limits of experimental accuracy, at most a one-degree rise, then the indicated positive ion current can be only 0.1 percent of the total interelectrode current.

SUMMARY

When stable electrical discharges are maintained between vacuum-insulated electrodes with gaps of $\frac{1}{8}$-inch and gap voltages of 80 to 100 kv, the negative particles decidedly predominate in the interelectrode current. If the current were composed of electrons and singly-charged positive ions, then the electrons outnumbered the positive ions by more than 300 to 1 for aluminum electrodes, and by more than 1000 to 1 for steel electrodes. The value for aluminum is in close agreement with that of Anderson,[3] and within the range of values obtained by Heard and Lauer[4] for copper. It also seems more than coincidental that previous results on secondary emission show that 500 to 1000 electrons are necessary for the emission of one positive ion. Since other work[1] has shown that fewer than 20 electrons are produced per positive ion under these conditions, it is evident that additional mechanisms such as field emission must assist in the release of this predominately negative emission.

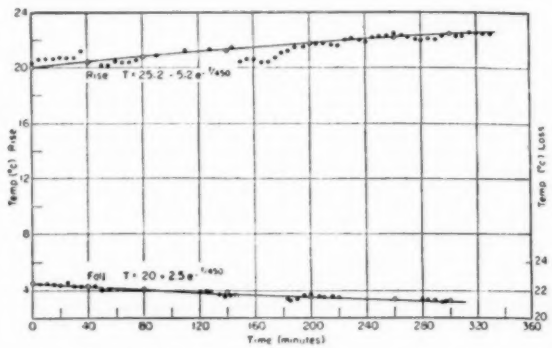


FIG. 3. Rising and falling temperature curves for aluminum cathode. Heating by positive particles of interelectrode current. Average total current = 10 μa and average total power = 0.90 w. Temperature perception = 0.2°C.

ACKNOWLEDGMENT

The author particularly wishes to acknowledge the assistance of Dr. John G. Trump of the High Voltage Laboratory at the Massachusetts Institute of Technology.

[3] Anderson, Doctorate thesis, Massachusetts Institute of Technology (1935).

[4] Heard and Lauer, University of California Radiation Laboratory Report No. 1622 (January, 1952).

Interaction of Moving Charges with Wave Circuits

J. R. PIERCE
Bell Telephone Laboratories, Inc., Murray Hill, New Jersey
(Received September 23, 1954)

When a charge moves along a nondispersive transmission line it induces forward and backward wave components in the line. For unaccelerated motion there is no radiation unless the charge moves with the speed of an unforced wave. For small accelerations, both radiation to the line and electromagnetic inertial effects caused by changes in the energy of the fields will be observed. The electromagnetic mass at velocities greater than the wave velocity is negative. Large accelerations can be handled by numerical computation. In the case of dispersive circuits, the charge radiates at frequencies for which the phase velocity of the circuit is equal to the velocity of the charge. This radiation is identified with Cherenkov radiation. Similar radiation takes place when a charge moves through a plasma.

INTRODUCTION

THE material presented here deals with the interaction of moving charges with an environment, called a circuit. The problem is attacked by methods commonly used in connection with microwave tubes, but the behavior disclosed, and the methods themselves, are related to phenomena and methods of classical electricity and magnetism.[1–3] This is apparent in connection with electromagnetic mass, radiation from an accelerated charge, and Cherenkov radiation.

In Part I of this paper, a very simple matter is considered: the motion of a charge near a nondispersive circuit. Matters dealt with include continuous transfer of energy from charge to circuit in the special case of synchronous velocity and electric forces on accelerated charges. Slow accelerations are dealt with by means of an electromagnetic mass, while rapid accelerations require another approach.

In Part II, uniform motion of a charge near a dispersive circuit is considered. In this case the feature of interest is the continuous transfer of energy to the circuit at any frequency for which the velocity of the charge is equal to the phase velocity of a mode of the circuit. This is a form of Cherenkov radiation, and Cherenkov radiation into a dielectric-filled wave guide is compared with Cherenkov radiation into an infinite dielectric medium.

The writer originally intended to include in this paper the theory of Cherenkov-like radiation to a smoothed-out electron gas (or plasma) by an electron moving through it. The treatment of this case follows from that of Part II; one needs merely to compute an impedance looking out into the gas. Because this matter has been treated in a somewhat different way by Pines and Bohm,[4] and because it is somewhat foreign to the body of the paper, this work has not been included. The result will be given, however. The power flow P from a charge q to a medium with plasma frequency ω_p is, in mks units,

$$P=\frac{K_0(\omega_p a/v)q^2\omega_p}{8\pi\epsilon a K_1(\omega_p a/v)}.$$

Here K_0 and K_1 are modified Bessel functions, and a is a lower-limiting or meaningful radius dependent on the spacing between the charges in the gas.

PART I

A MOVING CHARGE INTERACTING WITH A NONDISPERSIVE TRANSMISSION LINE

In this section we deal with the motion of a charge near a uniform, lossless, nondispersive transmission line. In dealing with a lossless, nondispersive line we can use very elementary mathematical methods. The procedure enables us to proceed by a series of small, clear steps and to keep our thoughts straight in doing so.

1. Differential Equation Approach

We will consider a uniform transmission line having a series inductance L and a shunt capacitance C per meter, as shown in Fig. 1. It will be assumed that a line charge density

$$q(z-vt) \text{ coulombs/meter}$$

moves parallel to and very close to the line all at the same velocity v, so that all the lines of force from portions of the charge terminate on the line at the same z-position at which they originate. Then the charge $q(z-vt)$ can be regarded as a charge induced on the line.

It will be assumed that all wave and particle velocities involved are small compared with the velocity of light.

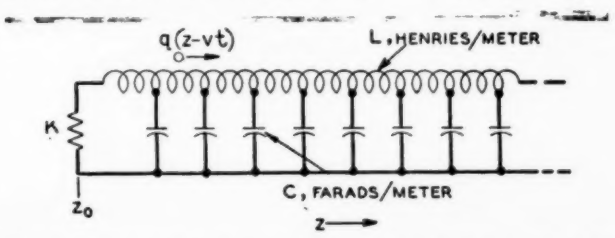


FIG. 1. A charge $q(z-vt)$ moves close to a distributed circuit which can carry an electromagnetic wave.

[1] G. A. Schott, *Electromagnetic Radiation* (The Cambridge University Press, New York, 1912).
[2] W. Heitler, *Quantum Theory of Radiation* (Clarendon Press, Oxford, England, 1944), second edition, Sec. I.
[3] L. I. Schiff, *Quantum Mechanics* (McGraw-Hill Book Company, Inc., New York, 1949), pp. 261–265.
[4] David Pines and David Bohm, Phys. Rev. **85**, 338–353 (1952).

Thus, if V is the voltage across the shunt capacitance, to a good approximation the field E in the z-direction at the inductance and at the charge is

$$E=-\partial V/\partial z. \tag{1.1}$$

The transmission line equations are then

$$\frac{\partial V}{\partial z}=-L\frac{\partial I}{\partial t}, \tag{1.2}$$

$$\frac{\partial I}{\partial z}=-C\frac{\partial V}{\partial t}+\frac{\partial}{\partial t}q(z-vt). \tag{1.3}$$

By differentiating and combining one obtains

$$\frac{\partial^2 V}{\partial z^2}-\frac{1}{c^2}\frac{\partial^2 V}{\partial t^2}=-L\frac{\partial^2}{\partial t^2}q(z-vt), \tag{1.4}$$

$$c=\frac{1}{(LC)^{\frac{1}{2}}}. \tag{1.5}$$

Here, c is the velocity of an unforced wave on the line.

If we assume the velocity v to be constant, we easily see that Eq. (1.4) has a solution made up of two parts. One is a special solution

$$V=-\frac{Lv^2}{1-(v/c)^2}q(z-vt). \tag{1.6}$$

It is convenient to rewrite this in terms of the characteristic impedance of the line, K

$$K=(L/C)^{\frac{1}{2}} \tag{1.7}$$

so that

$$V=-\frac{(v/c)}{1-(v/c)^2}Kvq(z-vt). \tag{1.8}$$

Here, $vq(z-vt)$ is the convection current of the electron stream. Thus, the voltage is given by the product of a current and an impedance times a dimensionless factor, which is as it should be.

To give a complete description of all possible excitations on the line we may add two unforced waves traveling to the left and right, solutions which make the left-hand side of Eq. (1.4) equal to zero. These are

$$V=f(z+ct), \tag{1.9}$$

$$V=g(z-ct), \tag{1.10}$$

where $f(z+ct)$ and $g(z-ct)$ may be arbitrary functions of the variable.

2. Summing Up Currents Induced in the Line

It seems of some interest and value to approach the problem in a slightly different manner. In this case, in deducing the fields produced by the moving charge, we will consider the charge as causing a current $J(z_1,t_1)$ amperes/meter to flow into the line, as shown in Fig. 2. Half of this current will flow to the right with a velocity c and half will flow to the left with the same velocity. We will then evaluate the current at a position z at a time t.

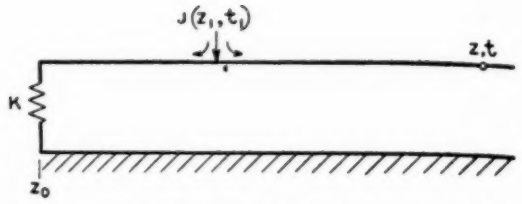


FIG. 2. A current density $J(z_1,t_1)$ flows into a transmission line and excites waves traveling to the left and to the right.

In order to make the problem reasonable physically, we will assume that the line extends indefinitely in the $+z$-direction, and is terminated in its characteristic impedance K at $z=z_0$.

We will call the part of the current at z at the time t due to waves traveling in the $+z$-direction I_+. This current may be expressed as

$$I_+=\tfrac{1}{2}\int_{z_0}^{z}J[z_1,t_1(z_1)]dz_1. \tag{2.1}$$

That is, it is a summing up of current which entered the line earlier and traveled to the right with the velocity c to reach point z at a time t. Thus

$$t_1(z_1)=t-\frac{z-z_1}{c}. \tag{2.2}$$

Now, if $q(z-vt)$ is the induced charge per meter, the induced current per meter, $J(z,t)$ is

$$\begin{aligned}J(z,t)&=\frac{\partial}{\partial t}q(z-vt)\\&=-vq'(z-vt),\end{aligned} \tag{2.3}$$

where $q'(x)$ is the derivative of the function with respect to its argument.

By using Eqs. (2.3) and (2.2) in connection with Eq. (2.3) we obtain

$$I_+=-\frac{v}{2}\int_{z_0}^{z}q'[(1-v/c)z_1-v(t-z/c)]dz_1. \tag{2.4}$$

This of course integrates directly to give

$$I_+=-\frac{v}{2(1-v/c)}\Big|_{z_0}^{z}q[(1-v/c)z_1-v(t-z/c)],$$

$$\begin{aligned}I_+=-\frac{v}{2(1-v/c)}\{&q(z-vt)\\&-q[(v/c)(z-ct)+z_0(1-v/c)]\}.\end{aligned} \tag{2.5}$$

We can obtain the part of the current at z_1 at a time t_1 due to current flowing to the left by a similar procedure

$$I_-=\tfrac{1}{2}\int_z^\infty J[z_1,t_1(z_1)]dz, \tag{2.6}$$

$$I_-=\tfrac{1}{2}\int_z^\infty q'[z_1(1+v/c)-c(t+z/c)]dz_1, \tag{2.7}$$

$$I_-=-\frac{vq(z-vt)}{2(1+v/c)}. \tag{2.8a}$$

Let us now examine Eqs. (2.5) and (2.8a). We see that I_+ consists of two parts. The first term in the brackets represents a current traveling along with the velocity v of the charge. The second term represents a current traveling along with the velocity of propagation c of the transmission system.

Suppose that we assume $v\neq c$, and consider the field in the vicinity of the charge far to the right of $z=z_0$, a long time after the charge has passed that point. If $q(z-vt)$ is a narrow function of z, that is, a pulse, the part of I_+ which travels with the velocity c will then be far away from the charge, and in the vicinity of the charge we will have

$$I_+=-\frac{v}{2(1-v/c)}q(z-vt) \tag{2.9}$$

and

$$I_-=-\frac{v}{2(1+v/c)}q(z-vt). \tag{2.8b}$$

The total current I will be the sum of I_+ and I_-, or

$$I=I_++I_-,$$

$$I=-\frac{1}{1-(v/c)^2}vq(z-vt). \tag{2.10}$$

We will note that I_+ and I_- can be considered as current traveling to the left and right as unforced waves with the velocity c. This is clear from Eqs. (2.1) and (2.6) together with Eq. (2.2). The charge in its motion continually builds up the current toward one edge of the charge and cancels it out toward the other edge, so that the *pattern* of current moves with the velocity v. Nonetheless the actual current flow at any instant can be considered as an excitation of the unforced waves of the transmission line, and if the charge suddenly vanished these waves would travel off to the left and to the right with their original spatial distribution but with a velocity c.

Thus, we can obtain the voltages V_+ and V_- corresponding to the currents (2.9) and (2.8b) by means of the usual transmission-line relations

$$V_+=KI_+, \tag{2.11}$$

$$V_-=-KI_-. \tag{2.12}$$

Here, Eq. (2.12) takes the form which it does because we have counted I_- as positive when it flows in the $+z$-direction.

Accordingly, we have

$$V_+=-\frac{1}{2(1-v/c)}Kvq(z-vt), \tag{2.13}$$

$$V_-=\frac{1}{2(1+v/c)}Kvq(z-vt), \tag{2.14}$$

$$V=V_++V_-,$$

$$V=-\frac{(v/c)}{1-(v/c)^2}Kvq(z-vt). \tag{2.15}$$

We see that this agrees with Eq. (1.8), the special solution of the differential equation.

The advance has been in breaking V and I up into $+$ and $-$ components, each simpler in form than their sum, and each obeying the usual transmission-line relations (2.11) and (2.12).

Before passing on, it is perhaps worth examining the expressions

$$I=-\frac{1}{1-(v/c)^2}vq(z-vt), \tag{2.10}$$

$$V=-\frac{(v/c)}{1-(v/c)^2}Kvq(z-vt). \tag{2.15}$$

As v approaches 0 the current I approaches the convection current of the moving charge. The voltage V, however, approaches zero.

For $v<c$ the voltage is proportional to the charge by a negative factor; that is, a positive charge produces a negative voltage (as in the case of a negative capacitance). In fact, at a given velocity the voltage has the same spatial distribution as the charge, and the relation is just as if the circuit consisted of a capacitance

$$-\frac{1-(v/c)^2}{(v/c)Kv}$$

farads per meter.

As v approaches c, that is, as the charge approaches synchronism with the unforced wave, I and V approach infinity. This is because I_+ and V_+ approach infinity, for I_- and V_- remain finite.

3. Charge Injected with Velocity of Unforced Water

The fact that the I and V of the special solution, as given by Eqs. (2.10) and (2.15), approach infinity as v approaches c does not mean that a charge shot along the line with the velocity c will produce an infinite field. In obtaining the expression (2.9) for I_+ it was assumed that a wave traveling with the velocity v became separated

from a wave traveling with the velocity c and this will not occur if $v=c$.

Let us investigate what does happen when $v=c$. In doing this, we will expand the second term in Eq. (2.5) in powers of $(1-v/c)$

$$\begin{aligned} &q[(v/c)(z-ct)+z_0(1-v/c)] \\ &\quad =q[(z-vt)-(z-z_0)(1-v/c)] \\ &\quad =q(z-vt)-(z-z_0)(1-v/c)q'(z-vt)+\cdots. \end{aligned} \tag{3.1}$$

Thus, for $v=c$, I_+ becomes

$$I_+=-\frac{v(z-z_0)}{2}q'(z-vt). \tag{3.2}$$

In other words, I_+ increases in amplitude as the distance from the beginning of the line, which is also the point of injection of the charge. Energy is continually transferred from the charge to the circuit, and the charge must experience a retarding force. We note that for a nondispersive circuit, an unaccelerated charge experiences a retarding force only when its velocity is equal to that for a wave on the circuit.

We note from Eq. (2.8) that there is no such peculiarity of behavior for the backward current, because the denominator contains the factor $(1+v/c)$ rather than $(1-v/c)$.

4. Changes in Velocity

Let us now consider cases in which $v\neq c$ and in which Eqs. (2.9), (2.8), (2.13), and (2.14) apply. Let us concern ourselves first with I_+

$$I_+=-\frac{vq(z-vt)}{2(1-v/c)}. \tag{2.9}$$

Now, suppose that at the time t_a when the charge is at z_a we suddenly change the velocity of the charge to v_a;

$$v_a=v+\delta v. \tag{4.1}$$

At the moment of change the current and field in the line cannot suddenly change. This will be true if the current of the new forward (+) waves [and that of the new backward (−) waves as well] does not change. The new I_+ must be made up of a component traveling with a velocity v_a [as in Eq. (2.9)] and one with a velocity c. The combination which is equal to Eq. (2.9) at $t=t_a$ and for which the part with velocity v_a obeys relation (2.9) is

$$\begin{aligned} I_+=&-\frac{v_aq[z-v_at+(v_a-v)t_a]}{2(1-v_a/c)} \\ &+\frac{v_aq[z-ct+(c-v)t_a]}{2(1-v_a/c)}-\frac{vq[z-ct+(c-v)t_a]}{2(1-v/c)}. \end{aligned} \tag{4.2}$$

We see that at $t=t_a$ this gives the same value as the previous expression for I_+.

We see that Eq. (4.2) consists of two parts; a part traveling with the speed of the charge, and a part traveling with the velocity of propagation of the line, c. If we assume that δv in Eq. (4.2) is very small, we can rewrite these to give

$$\begin{aligned} I_+=&-\left[v+\frac{\delta v}{(1-v/c)}\right]\frac{q[z-(v+\delta v)t+\delta vt_a]}{2(1-v/c)} \\ &+\frac{\delta v}{2(1-v/c)^2}q[z-ct+(c-v)t_a]. \end{aligned} \tag{4.3}$$

Thus, the amplitude of the current accompanying the charge is changed, and a small part of the current proportional to the velocity change δv is shaken off or radiated and travels as an unforced wave with the velocity c. As time passes, the pulses represented by the two components will separate, forming two distinct and essentially nonoverlapping pulses.

It is now of interest to examine the energy carried by these pulses. In terms of the inductance L per unit length and the capacitance C per unit length, the total stored magnetic and electric energy per unit length W is

$$W=\tfrac{1}{2}(I^2L+V^2C). \tag{4.4}$$

This can be expressed in terms of I_+ and I_- as

$$W=W_++W_-=\tfrac{1}{2}[(I_++I_-)^2L+(I_+-I_-)^2K^2C]. \tag{4.5}$$

Here, W_+ is the stored energy per unit length, the forward wave and W_- is the stored energy per unit length of the backward wave. By using Eqs. (1.5) and (1.7) we obtain

$$W_+=\frac{K}{c}I_+^2; \quad W_-=\frac{K}{c}I_-^2. \tag{4.6}$$

To get the total energy associated with the forward and backward waves we should integrate W_+ and W_- as given by Eq. (4.6) with respect to z at a particular time. The only term that varies with distance is the square of the charge, and we will call the integral of the square of the charge per unit length with respect to distance Q^2/L. Then the energy corresponding to the first part of Eq. (4.3), the pulse moving with the charge, which we will call W_{+M}, is

$$W_{+M}=\frac{(v/c)^2Kc}{4(1-v/c)^2}\frac{Q^2}{L}\left[1+\frac{2(\delta v/v)}{(1-v/c)}+\frac{(\delta v/v)^2}{(1-v/c)^2}\right]. \tag{4.7}$$

While the part corresponding to the second part of Eq. (4.3), which we will call W_R will be

$$W_{+R}=\frac{(v/c)^2KcQ^2(\delta v/v)^2}{4(1-v/c)^4L}. \tag{4.8}$$

The forward wave part of the electromagnetic energy carried along with the charge is W_{+M} while W_{+R} is the forward wave part of the electromagnetic energy detached from the charge or "radiated."

Suppose we reach the final velocity v_a in two discrete steps, carried out at times far enough separated so that the "radiated" pulses do not overlap. Then the total "radiated" energy will be halved, while to the first order the change in the energy carried along will be the same as before.

We can apply this argument only to sudden changes in velocity separated in time enough so that the "radiated" pulses do not overlap. However, it indicates very strongly that for low accelerations the electromagnetic energy carried along becomes independent of acceleration while the energy radiated becomes negligible.

To this approximation, for small accelerations we will neglect the radiated energy and take the energy carried along by the forward component

$$W_{+M}=\frac{(v/c)^2KcQ^2}{4(1-v/c)^2L}. \tag{4.9}$$

If we make further use of these methods we find the energy carried along by the backward component to be

$$W_{-M}=\frac{(v/c)^2KcQ^2}{4(1+v/c)^2L}. \tag{4.10}$$

Thus, W_M the total energy carried along is

$$W_M=W_{+M}+W_{-M},$$

$$W_M=\frac{(v/c)^2[1+(v/c)^2]KcQ^2}{2[1-(v/c)^2]^2}. \tag{4.11}$$

Now, suppose that we assume a law of force

$$f=dM/dt, \tag{4.12}$$

where f is the force and M is momentum. We know that

$$dW/dt=fv=\frac{dM}{dt}v, \tag{4.13}$$

$$M=\int dW/v. \tag{4.14}$$

Or, integrating by parts

$$M=(W/v)+\int Wdv/v^2. \tag{4.15}$$

It is easiest to divide the contributions to momentum into two parts, that M_{+M} from Eq. (4.9) and M_{-M} from Eq. (4.10). This gives

$$M_{+M}=\frac{KQ^2}{4L}\frac{1}{(1-v/c)^2}, \tag{4.16}$$

$$M_{-M}=-\frac{KQ^2}{4L}\frac{1}{(1+v/c)^2}. \tag{4.17}$$

Whence

$$M=M_{+M}+M_{-M},$$

$$M=\frac{KQ}{L}\frac{(v/c)}{[1-(v/c)^2]^2}. \tag{4.18}$$

We note that this is an odd function of the velocity, as it should be.

5. The Force on the Charge

In Sec. 4 an electromagnetic momentum M was deduced from a definition of momentum, relation (4.12), and a relation between energy change and work, relation (4.13).

Physically, we are dealing with a charge moving in certain fields. Any "electromagnetic momentum" must manifest itself as an electric field which acts in the direction of motion when the speed of the charge is changed.

Let us, then, consider the fields present in the vicinity of a charge after the velocity of the charge has been changed by an amount δv. Let us deal first with the fields of the wave components traveling in the $+z$-direction, as expressed by Eq. (4.3). We have

$$V_+=KI_+, \tag{2.11}$$

$$E_+=-\frac{\partial}{\partial z}(KI_+) \tag{5.1}$$

and from Eq. (4.3)

$$E_+=K\left(v+\frac{\delta v}{(1-v/c)}\right)\frac{q'[z-v_at+(v_a-v)t_a]}{2(1-v/c)}$$

$$-K\frac{\delta v}{2(1-v/c)^2}q'[(z-vt)+(c-v)(t_a-t)]. \tag{5.2}$$

Now, the impulse H_+ given to the charge due to the change in velocity will be given by the integral

$$H_+=\int_{t=t_a}^{\infty}\int_{z=-\infty}^{\infty}E_+q[z-v_at+(v_a-v)t_a]dzdt. \tag{5.3}$$

Here we have multiplied the field by an element of charge and integrated over all the charge, and then we have integrated this force with respect to time in order to obtain the impulse.

From Eq. (5.2) we see that there will be two terms in the integral. The first gives zero

$$\int_{t=t_a}^{\infty}\int_{z=-\infty}^{\infty}q'[z-v_at+(v_a-v)t_a]q[z-v_at+(v_a-v)t_a]dzdt$$

$$=\int_{t-t_a}^{\infty}dt\Big|_{z=-\infty}^{\infty}\frac{q^2[z-v_at+(v_a-v)t_a]}{2}=0. \tag{5.4}$$

The second term has a factor

$$\int_{t=t_a}^{\infty}\int_{z=-\infty}^{\infty} q'[(z-vt)+(c-v)(t_a-t)] \times q[z-v_at+(v_a-v)t_a]dzdt. \quad (5.5)$$

This can be integrated by the following substitution of variables:

$$t=u_1, \quad (5.6)$$

$$z-v_at+(v_a-v)t_a=u_2. \quad (5.7)$$

The element of area is obtained by multiplying du_1du_2 by the Jacobian

$$\begin{vmatrix} \frac{\partial u_1}{\partial t} & \frac{\partial u_1}{\partial z} \\ \frac{\partial u_2}{\partial t} & \frac{\partial u_2}{\partial z} \end{vmatrix} du_1du_2=\begin{vmatrix} 1 & 0 \\ -v & 1 \end{vmatrix} du_1du_2=du_1du_2. \quad (5.8)$$

The integral then becomes

$$\int_{u_1=t_a}^{\infty}\int_{u_2=-\infty}^{\infty} q'[u_2+(c-v)(t_a-u_1)]q(u_2)du_2du_1. \quad (5.9)$$

We integrate first with respect to u_1, and obtain

$$\int_{u_2=-\infty}^{\infty} q(u_2)du_2\Big|_{u_1=t_a}^{\infty} - \frac{q[u_2+(c-v)(t_a-u_1)]}{(c-v)}$$

$$=\frac{1}{(c-v)}\int_{u_2=-\infty}^{\infty}[q(u_2)]^2du_2=\frac{Q^2}{(c-v)L}. \quad (5.10)$$

We note that we have used the symbol Q^2/L for the integral before.

We have now evaluated the integral occurring in Eq. (5.3) and from Eqs. (5.2), (5.3), and from this work we see that

$$H_+=-\frac{KQ^2\delta v}{2c(1-v/c)^3L}. \quad (5.11)$$

This H is the integral of the force experienced by the charge following a velocity change δv, times the time. If we express it as the result of a change in a momentum M, the relation should be of the form

$$H=\int Fdt$$

$$F=-dM/dt,$$

$$\int Fdt=H=-\delta M,$$

or

$$\frac{\delta M}{\delta v}=\frac{dM}{dv}=-\frac{KQ^2}{2c(1-v/c)^3L}. \quad (5.12)$$

If we compare this with Eq. (4.16) we see that Eqs. (5.12) and (4.16) differentiated are in agreement. In Eq. (4.14) the momentum was defined in terms of change with velocity of stored electromagnetic energy. We have seen that the force which resists (or aids) slow acceleration of the charge is actually the force exerted on the charge by the field of the wave which is radiated when the charge is accelerated. Although the energy radiated approaches zero as the acceleration between two particular velocities is made slower and slower, the impulse due to the radiated field does not approach zero but approaches a constant value instead. This is because the energy depends on the square of the radiated field while the impulse depends on the first power of the radiated field.

In other words, for a given change in velocity, for the radiated field the integral $\int_{-\infty}^{\infty} E^2dz$ approaches zero as the acceleration is made slow, but the integral $\int_{-\infty}^{\infty} |E|dz$ approaches a constant which is not zero.

Because the impulse depends on the first power of the radiated field, the calculations which have been made concerning it hold even when pulses representing the radiated field overlap. Hence, we may regard the momentum M as correct for all accelerations which are gradual enough so that the velocity does not change substantially in the time it takes the radiated pulse due to the velocity change to pass the charge.

6. Velocities Greater Than the Velocity of the Natural Wave

The queer feature about the total momentum as defined by Eq. (4.18) is that for $v>c$, $\partial M/\partial v<0$. Thus, if the velocity of the charge in the $+z$-direction is increased, the charge experiences a force in the $+z$-direction. Contrarily, if the velocity of the charge in the $+z$-direction is reduced, the particle experiences a retarding force. Thus, motion with a uniform velocity greater than c would seem to be unstable; it seems that the charge would tend either to speed up or to slow down. Physical considerations would suggest that it might slow down.

Suppose that the charge has a mechanical mass m. If we assume its acceleration to be small, the following equation should be satisfied

$$(dM/dt)+(mdv/dt)=0, \quad (6.1)$$

$$M+mv=\text{const.} \quad (6.2)$$

Thus, the charge cannot spontaneously change its velocity with a small acceleration. This need not embarrass us, for Eq. (6.1) holds for small accelerations only.

In principle, motion under large accelerations can be computed numerically. Suppose that an abrupt change is made in the velocity of the charge. The radiated fields acting during a short interval following the change can be computed using Eqs. (4.2) or (4.3), and an impulse can be obtained by integrating charge times mass over

the interval. At the end of the first short interval another abrupt change in velocity, equal to the impulse divided by the mass, can be assumed, and new radiated fields can be computed. Over the next interval the radiated fields of the first two intervals act on the charge. The calculation can be continued thus, step by step. Alternatively, the equation of motion can be written in the form of an integral equation.

It is interesting to note that because of the electromagnetic mass we cannot gradually accelerate a charge past the velocity c, but this does not forbid shooting it into the system with a velocity greater than c, or even rapidly accelerating it to a velocity greater than c.

It seems likely that even when $v>c$ and $|\partial M/\partial v|$ is smaller than m, the motion will be stable. An approximate calculation was made for $v>c$ and for $|\partial M/\partial v|>m$. It showed a rapid deceleration of the moving charge. On the other hand, a calculation for $v>c$ and $|\partial M/\partial v|<m$ showed the charge settling down to a constant velocity.

If a charge is shot into a line with a velocity $v>c$, the field corresponding to the second term of Eq. (2.5) will cause an initial deceleration of the charge, and the fields radiated subsequently will decelerate the charge further.

Thus, a charge shot into a line with a speed $v>c$ tends to slow down and to radiate *provided that* $|\partial M/\partial v|$ *is greater than the mechanical mass* m.

7. Charge Separated from the Circuit

So far our picture has been much like that of a physically idealized transmission line with a charge very close to it. What happens if we separate the charge from the line? In the case of slow waves we can use an electrostatic approach in which an element of charge is regarded as exciting the line over a distance, with some appropriate weighting. Perhaps a better approach is to regard the actual transmission system as having many modes, active and passive. From this point of view, Part I can be regarded as having dealt with one mode of transmission of a realistic multimode transmission system.

PART II

UNACCELERATED MOTION WITH A DISPERSIVE CIRCUIT

We have seen that when a charge moves with a constant velocity near a lossless-nondispersive circuit it gives up energy continually only when the velocity of the charge is equal to the wave velocity of the circuit. For dispersive circuits the phase velocity varies with frequency and thus over a wide range of charge velocity a wave of a particular frequency will have the charge velocity. This causes a continual transfer of energy from the charge to the circuit. Cherenkov radiation is an example of this. We will consider other simpler examples first and then discuss Cherenkov radiation.

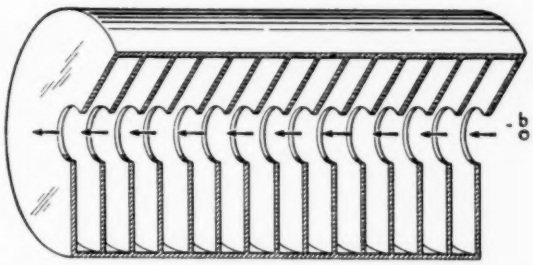


FIG. 3. A charge q flows through a series of resonators.

8. A Charge Passing Through a Series of Resonators

Imagine a charge q to travel with a velocity v through apertures in a series of thin, pillbox resonators, as shown in Fig. 3. Imagine that the holes are very small in diameter and that the resonators are so thin that the charge passes across one in a time small compared with a cycle of the natural frequency of a resonator.

Then, the passage of the charge across a given resonator corresponds to transferring a charge from one side of the resonator to the other, and if C_R is the effective capacitance of a given mode of the resonator referred to the point at which the electron passes through, an energy $W=q^2/2C_R$ is transferred to that mode of the resonator. The resonator will oscillate with this energy after the particle has passed. The resonators which the charge has passed oscillate independently, and they are phased so that the oscillations constitute a wave of phase velocity equal to the charge velocity v. The group velocity of the wave is of course zero; there is no energy flow from resonator to resonator.

If d is the distance between centers of resonators, the power P flowing from the electrons to the resonators is[5]

$$P=\frac{Wv}{d}\frac{q^2v}{2dC_R}. \tag{8.1}$$

As we make the resonators thinner and thinner, C_R tends to be proportional to $1/d$ and hence we may regard dC_R as a constant which does not depend on d.

Let us consider the retarding force F against which the electron works:

$$P=Fv=q^2v/2dC_R, \tag{8.2}$$

$$F=q^2/2dC_R. \tag{8.3}$$

Thus, the force is independent of velocity and proportional to the square of the charge.

9. Use of Fourier Transforms

It was possible to treat the case in Sec. 1 by very elementary means. In more complicated problems it is convenient to represent the charge by means of its Fourier transform. Let the charge per unit length q have

[5] Here we disregard all but one mode of oscillation of the resonator. Other modes will give other contributions to the power.

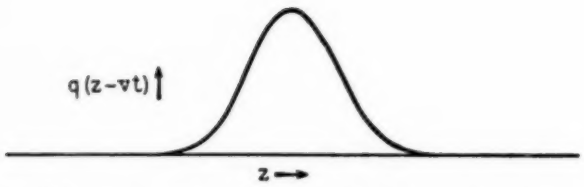


FIG. 4. The spatial distribution of the charge $q(z-vt)$.

some spatial distribution $q(z-vt)$ such that

$$\int_{-\infty}^{\infty} q(z-vt)dz=q. \tag{9.1}$$

Here t is regarded as a constant in the integration. We can also represent the charge as

$$q(z-vt)=\int_{-\infty}^{\infty} g(\beta)e^{-j\beta(z-vt)}d\beta. \tag{9.2}$$

The function $g(\beta)$ is given by

$$g(\beta)=\frac{1}{2\pi}\int_{-\infty}^{\infty} q(z-vt)e^{j\beta(z-vt)}dz. \tag{9.3}$$

Here we should note what sort of function of $\beta g(\beta)$ is. We will consider $q(z-vt)$ to represent a relatively narrow lump of charge, as shown in Fig. 4. From Eq. (9.3) we see that for small values of β,

$$e^{j\beta(z-vt)}$$

will be constant and nearly unity for all values of $(z-vt)$ for which $q(z-vt)$ is substantially different from zero. Thus, from Eq. (9.1) we see that for small values of β

$$g(\beta)=q/2\pi \tag{9.4}$$

and for larger values of β, $g(\beta)$ will decrease in value, perhaps as shown in Fig. 5.

From Eq. (9.2) we can obtain the current I in the z-direction either by multiplying the charge by v, or by using the relation

$$\partial I/\partial z=-\partial q/\partial t, \tag{9.5}$$

$$I=v\int_{-\infty}^{\infty} g(\beta)e^{-j\beta(z-vt)}d\beta. \tag{9.6}$$

In connection with the current expressed as in Eq. (9.6) we can obtain the field produced by the current through the use of an impedance per unit length $Z(\omega,\beta)$.

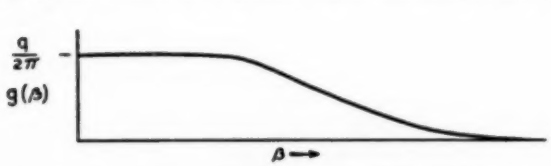


FIG. 5. The Fourier transform $g(\beta)$ of the charge $q(z-vt)$.

In terms of this impedance per unit length

$$E=-v\int_{-\infty}^{\infty} Z(\omega,\beta)g(\beta)e^{-j\beta(z-vt)}d\beta. \tag{9.7}$$

Here, we see that

$$\omega=v\beta. \tag{9.8}$$

We should note that for circuits of finite transverse extent involving no loss (resistance, conductance) Z will be purely imaginary. However, for actual circuits, which always have some loss, Z may be nearly imaginary over most of the range of ω and β but whatever real part Z has, that real part will be positive. This is important in connection with the poles of the idealized purely imaginary (reactive) impedances considered later.

The instantantaneous power flow *from the electrons to the field* (circuit) is

$$P=-\int_{-\infty}^{\infty} EIdz. \tag{9.9}$$

Now

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\gamma)e^{j\gamma z}h(\beta)e^{j\beta z}d\gamma d\beta dz = 2\pi\int_{-\infty}^{\infty} f(\beta)h^*(\beta)d\beta. \tag{9.10}$$

Whence, from Eqs. (9.6) and (9.7)

$$P=2\pi v^2\int_{-\infty}^{\infty} Z(\omega,\beta)g(\beta)g^*(\beta)d\beta. \tag{9.11}$$

10. Problem of Section 8 by Fourier Transform Method

Let us consider the problem of Sec. 8, disregarding all but one mode of the resonators. The impedance of one resonator Z_R will be

$$Z_R=\frac{-j}{C_R[\omega-(\omega_0^2/\omega)]}, \tag{10.1}$$

$$\omega_0^2=LC. \tag{10.2}$$

The impedance per unit length, Z, will be

$$Z=\frac{Z_R}{d}=\frac{-j\omega}{dC_R(\omega^2-\omega_0^2)}. \tag{10.3}$$

We remember that dC_R is independent of the resonator spacing d.

According to Eqs. (9.11) and (9.8), the power flow to the circuit will be

$$P=2\pi v^2\int_{-\infty}^{\infty}\left[\frac{-j\beta}{dC_R v[\beta-(\omega_0/v)][\beta+(\omega_0/v)]}\right] \times g(\beta)g^*(\beta)d\beta. \tag{10.4}$$

Now, because $q(z-vt)$ is real, $g(\beta)g^*(\beta)$ is an even function of β. The impedance (10.3), the bracketed part in the integrand, is an odd function of β with two poles, as shown in Fig. 6. Except for the contribution in passing the poles, the integral would be zero. How do we get by the poles?

We note that for an actual circuit the impedance will not pass through infinity but will assume a high, real positive value. Hence, ω_0 must [see Eq. (10.4)] have a small positive imaginary part which we will call $j\delta$. Near the pole at $\beta=\omega_0/v$, for instance, we can disregard the variation with respect to β except for that in the factor $\beta-\omega_0/v$ and write for the contribution to the integral from $\beta=\omega_0/v-\alpha$ to $\beta=\omega/v_0+\alpha$

$$\Delta P=\frac{-j\pi v g(\omega_0/v)g^*(\omega_0/v)}{dC_R}\times\int_{(\omega_0/v)-\alpha}^{(\omega_0/v)+\alpha}\frac{d\beta}{[\beta-(\omega_0/v)]-j\delta}, \tag{10.5}$$

$$\Delta P=\frac{-j\pi v g(\omega_0/v)g^*(\omega_0/v)}{dC_R}\ln\frac{\alpha-j\delta}{-\alpha-j\delta}. \tag{10.6}$$

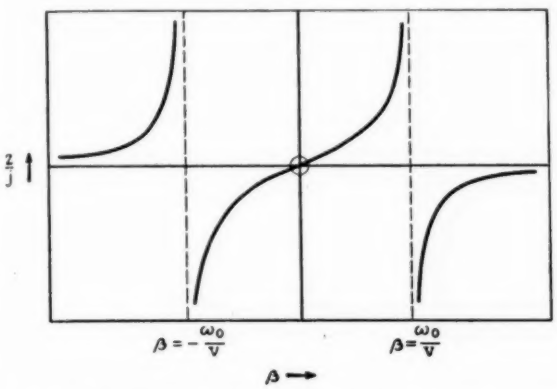


FIG. 6. The impedance Z of a circuit, plotted *vs* phase constant β, showing poles.

For $\alpha=0$ the logarithm is zero; for $\alpha\gg\delta$ it is $\pm j\pi$. We note that for intermediate values the argument has a positive imaginary part, and hence the value is $+j\pi$.

There is an equal contribution from the pole at $\beta=-\omega/v_0$, and hence the value of the total integral is

$$P=\frac{2\pi^2 v g(\omega_0/v)g^*(\omega_0/v)}{dC_R}. \tag{10.7}$$

If we assume that ω_0/v is small in the sense discussed in connection with Eq. (9.4), then Eq. (9.4) applies and

$$P=q^2v/2dC_R. \tag{10.8}$$

This agrees with Eq. (1.2), the assumption that ω_0/v is small merely saying, as we did in Sec. 9, that the charge passes through each resonator in a small part of a cycle of oscillation at the natural frequency of the resonator.

11. Nondispersive Line

Suppose that we have a transmission system consisting of a distributed series inductance L per unit length and distributed shunt capacitance C per unit length. Let us consider the interaction of the charge with the fields associated with the mode of propagation of this circuit. If we consider slow waves we can take the longitudinal field as $-\partial V/\partial z$. Under these circumstances the impedance Z per unit length is found to be

$$Z=\frac{-j\omega L\beta^2}{\beta^2-\omega^2LC}. \tag{11.1}$$

From Eqs. (2.8) and (2.11) we obtain

$$P=\frac{-j2\pi v^3L}{1-v^2LC}\int_{-\infty}^{\infty}\beta g(\beta)g^*(\beta)d\beta. \tag{11.2}$$

We see that the integrand is an odd function of β with no poles. Hence, the power is zero (except when $v^2LC=1$; this case must be treated differently, as, for instance, in Sec. 2.3).

12. Dispersive Circuits

When the circuit is dispersive and has for some frequency a phase velocity equal to the velocity of the charge, then there is a pole in the integrand of Eq. (9.11), as in the case of Eq. (10.4), and the power is not zero.

For instance, consider a wave guide filled with dielectric material, with a charge moving on the axis. There is an infinite number of transverse magnetic axially symmetrical modes of propagation which can be excited by the charge. Each is dispersive; the phase velocity varies from ∞ at a cutoff frequency to $1/(\mu\epsilon)^{\frac{1}{2}}$ at very high frequencies.

Thus, if the charge travels faster than $1/(\mu\epsilon)^{\frac{1}{2}}$, for each such mode of propagation there will be a pair of simple poles in the integral (9.11), and thus each mode will contribute something to the power given up by the moving charge.

Now, if we make the wave guide very large in diameter, the charge is to all intents in a large space of dielectric constant ϵ, and we should approach the usual case of Cherenkov radiation.

It seems worth while to examine this matter in detail and accordingly, the power flow from a moving charge distribution to *one axially symmetrical transverse magnetic* mode of a circular wave guide of radius a filled with a dielectric of dielectric constant ϵ will be evaluated. The relation between the field and the exciting current is[6]

$$E_z=\frac{-j(\Gamma^2+\beta_0^2)}{\omega\epsilon}e^{-\Gamma z}\Sigma_n\frac{\Pi_n(x,y)J_n}{\Gamma_n^2-\Gamma^2}, \tag{12.1}$$

[6] J. R. Pierce, *Traveling Wave Tubes* (D. Van Nostrand Company, Inc., New York, 1950), Chapter VI.

where

$$J_n=\frac{\int\int J(x,y)\Pi_n(x,y)dxdy}{\int\int[\Pi_n(x,y)]^2dxdy}. \tag{12.2}$$

We will take

$$\Pi_n(x,y)=J_0(\gamma_n r), \tag{12.3}$$

where

$$J_0(\gamma_n a)=0. \tag{12.4}$$

Then

$$\int\int[\Pi_n(x,y)]^2dxdy$$

$$=\int_0^a 2\pi r[J_0(\gamma_n r)]^2dr=\pi a^2[J_1(\gamma_n a)]^2. \tag{12.5}$$

Here, use has been made of Eq. (12.4).

We will assume a current I to flow at $r=0$, so that

$$\int\int J(x,y)\Pi_n(x,y)dxdy=1. \tag{12.6}$$

Whence

$$J_n=\frac{I}{\pi a^2[J_1(\gamma_n a)]^2}. \tag{12.7}$$

Thus, from the relation

$$E=-IZ$$

we deduce that the impedance Z_n associated with the nth mode is

$$Z_n=\frac{j(\Gamma^2+\beta_0^2)}{\pi a^2\omega\epsilon[J_1(\gamma_n a)]^2(\Gamma_n^2-\Gamma^2)}. \tag{12.8}$$

Here

$$\omega=\beta v, \tag{12.9}$$

$$\beta_0=\omega(\mu\epsilon)^{\frac{1}{2}}=\frac{\omega}{u}=\frac{v}{u}\beta, \tag{12.10}$$

$$\Gamma=j\beta. \tag{12.11}$$

And, for the wave guide considered

$$\Gamma_n^2=-\beta_n^2,$$
$$\beta_n^2+\gamma_n^2=\beta_0^2. \tag{12.12}$$

So that

$$Z_n=\frac{-j\beta}{\pi a^2\epsilon v[J_1(\gamma_n a)]^2}\times\left[\beta-\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}\right]\left[\beta+\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}\right]. \tag{12.13}$$

Thus, for this mode the power P_n will be

$$P_n=\frac{2v}{a^2\epsilon[J_1(\gamma_n a)]^2}\times\int_{-\infty}^{\infty}\frac{-j\beta g(\beta)g^*(\beta)d\beta}{\left[\beta-\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}\right]\left[\beta+\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}\right]}. \tag{12.14}$$

This is of exactly the same form as Eq. (10.4), and by the same process we obtain

$$P_n=\frac{2\pi v g\left(\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}\right)g^*\left(\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}\right)}{a^2\epsilon[J_1(\gamma_n a)]^2}. \tag{12.15}$$

If we assume that

$$\frac{\gamma_n}{[(v/u)^2-1]^{\frac{1}{2}}}$$

is sufficiently small so that we can use Eq. (9.4), we obtain

$$P_n=\frac{q^2v}{2\pi a^2\epsilon[J_1(\gamma_n a)]^2}. \tag{12.16}$$

To get the total power for all modes we would have to sum with respect to n.

Let us rather examine the power as a function of frequency. At a given point there is no radiation present in the wave guide prior to the passage of the charge. After the passage of the charge, each mode is excited at a particular frequency such that the phase velocity equals the velocity of the charge. Thus, the field has a line spectrum.

Let ω_n be the cutoff radian frequency of the nth mode. Then

$$\gamma_n=\omega_n(\mu\epsilon)^{\frac{1}{2}}=\frac{\omega_n}{u}. \tag{12.17}$$

The radian frequency ω_n' at which the phase velocity of the wave is equal to the velocity of the charge is such that

$$v(\mu\epsilon)^{\frac{1}{2}}=\frac{v}{u}=\left[1-\left(\frac{\omega_n}{\omega_n'}\right)^2\right]^{-\frac{1}{2}},$$
$$\frac{\omega_n'}{\omega_n}=\left[1-\left(\frac{u}{v}\right)^2\right]^{-\frac{1}{2}}. \tag{12.18}$$

Let us define

$$\Delta\omega=\omega_{n+1}'-\omega_n'. \tag{12.19}$$

Then

$$\Delta\omega=\frac{\omega_{n+1}-\omega_n}{[1-(u/v)^2]^{\frac{1}{2}}}. \tag{12.20}$$

For large values of n

$$\gamma_{n+1}a-\gamma_n a=\frac{\omega_{n+1}a}{u}-\frac{\omega_n a}{u}=\pi, \tag{12.21}$$

$$\omega_{n+1}-\omega_n=\pi u/a,$$

$$\Delta\omega=\frac{\pi u}{a[1-(u/v)^2]^{\frac{1}{2}}}, \tag{12.22}$$

$$\frac{P_n}{\Delta\omega}=\frac{q^2[(v/u)^2-1]^{\frac{1}{2}}}{2\pi^2 a\epsilon[J_1(\omega_n a u)]^2}.$$

Now, for large values of $\omega_n a/u$ at the extrema where

$$J_0\left(\frac{\omega_n a}{u}\right)=0, \tag{12.23}$$

$$\left[J_1\left(\frac{\omega_n a}{u}\right)\right]^2=\left(\frac{2u}{\pi\omega_n a}\right)^2,$$

hence

$$\frac{P_n}{\Delta\omega}=\frac{q^2\omega_n[(v/u)^2-1]^{\frac{1}{2}}}{4\pi\epsilon u}, \tag{12.24}$$

$$\frac{P_n}{\Delta\omega}=\frac{q^2\omega_n'[(v/u)^2-1]}{4\pi\epsilon u}.$$

This is in agreement with the power radiated per unit radian frequency in Cherenkov radiation.[2]

13. Comparison with Cherenkov Radiation

We have seen that the expression for the radiation per unit frequency of a charge moving along the axis of a large dielectric-filled guide is the same as the expression for the power of Cherenkov radiation. We feel that this is as it should be, and yet there are several matters which may appear puzzling.

First, we are inclined to regard an infinite, homogeneous, isotropic space as nondispersive, and yet we have seen that radiation into a wave guide depends on the dispersive nature of the modes of propagation.

As a matter of fact, an infinite homogeneous space will support all dispersive modes. It differs from a wave guide chiefly in having a continuous rather than a discrete spectrum of modes. If we wish, the characteristic pattern of Cherenkov radiation can be regarded as made up of a continuous spectrum of modes such as those used in Sec. 12.

The over-all pattern of Cherenkov radiation is simple. It is shown in Fig. 7. The pattern is characteristic of a shock wave. The wave front of the radiation is a vee extending back from the moving charge. The wave front advances with a speed u

$$u=\frac{1}{(\mu\epsilon)^{\frac{1}{2}}}$$

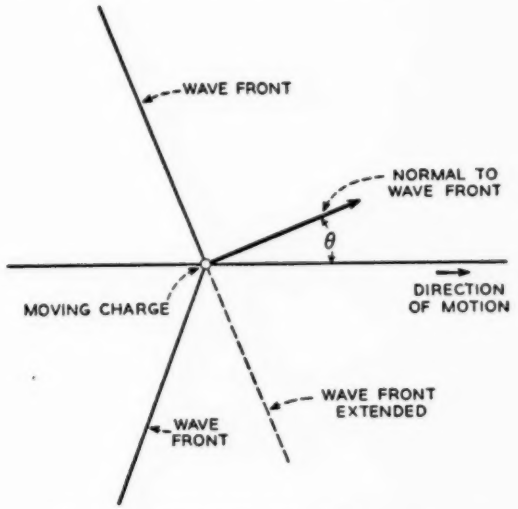


FIG. 7. The wave front of Cherenkov radiation.

normal to itself. The intersection of the wave front travels along the path of the particle with the particle speed v, which is greater than u. If θ is the angle between the normal to the wave front and the path of the particle, we see that

$$u=v\cos\theta. \tag{13.1}$$

The angle θ is the angle of the cone of radiation in Cherenkov radiation.

We see that the velocity of a *wave front* along a line may be greater than the velocity of a plane electromagnetic wave. In fact, the Cherenkov radiation clearly travels so as to make the velocity of the wave front along the path equal to the velocity of the moving charge.

We may note that the angle θ in Eq. (13.1) is characteristic of the resolution of wave-guide modes into sums of plane electromagnetic waves, in which case u is, as here, the velocity of a plane electromagnetic wave in the medium and v is the phase velocity of the wave in the guide.

Another thing which may at first seem puzzling is that Cherenkov radiation has a continuous spectrum, while the radiation in the wave guide has a line spectrum. The moving charge cannot be aware of the walls of the wave guide, because the charge is moving faster than the group velocity of any electromagnetic signal in the medium. We are unused to devices (in this case a wave guide) which seem to put all the power of a source with a continuous spectrum into a line spectrum.

This can be made clear by means of a physical picture. In order to make matters as simple as possible, let us consider a two-dimensional case, corresponding to a line charge moving between infinite parallel conducting planes, as shown in Fig. 8.

The reflection of the wave front from the conducting plane gives the lozenge-shaped wave front shown in Fig. 8. We may if we wish regard this pattern of wave

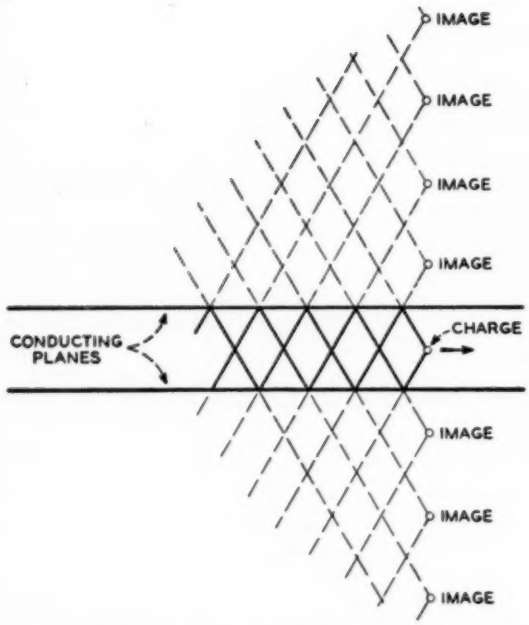


FIG. 8. Cherenkov radiation of a line charge between parallel planes—the wave fronts of the charge and of its images (dashed). The radiation has a line spectrum.

front as made up of the wave fronts of image charges, which are shown dotted outside of the planes in Fig. 8.

It is clear that the successive passages of the reflected wave fronts will result in a line spectrum rather than a continuous spectrum. However, the charge does not know that the reflecting walls are there. How, then does the presence of the walls convert the continuous spectrum of Cherenkov radiation into a line spectrum? Perhaps we should first note that in any case the charge itself sees a steady retarding field. We should also note that if we add the walls while the charge is moving, only the radiation which has not yet reached the position of the walls is reflected. The continuous spectrum of Cherenkov radiation can in principle be deduced only by an extended examination of the radiation pattern, while the line spectrum of the guide can be deduced only by an extended examination of the different radiation pattern in the guide. Now, one pattern which has been examined for a long period cannot be converted into the other.

There are other cases in which a signal which we think of as having a continuous spectrum can be converted into a signal which we think of as having a line spectrum. Let a very short pulse come along the transmission line of Fig. 9 from the left, and let the switch be closed after the pulse has passed but before the first reflection returns to the switch. All the electromagnetic energy of the pulse is now batting back and forth between the switch and the end of the line. The single pulse on a line, which we ordinarily think of as having a continuous spectrum, has been converted into a recurrent phenomenon with a line spectrum. This does not mean, however, that a linear, time-invariant, network can put all the energy of a signal source having a continuous spectrum into a number of discrete frequencies. All a linear network can do is reject or absorb some frequencies while passing others.

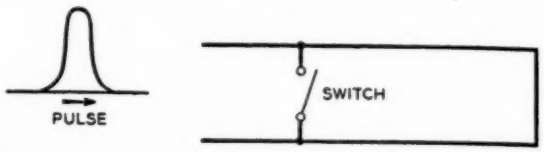


FIG. 9. A pulse trapped in a section of line assures a line spectrum.

A third question which might arise concerns the very fact of a charged particle traveling through a medium with a speed greater than that of an electromagnetic wave. In the case of Cherenkov radiation, this is possible because the particle is shot into the medium from outside. This does not show that it is possible to accelerate a particle within a homogeneous, isotropic dielectric to a speed greater than the speed of a plane electromagnetic wave in the medium, and indeed, if the medium is linear for all field levels this would seem to take an infinite energy under a variety of assumptions, provided that the acceleration is small.

ACKNOWLEDGMENT

The writer wishes to acknowledge the valuable comments and criticisms of Dr. P. L. Hartman.

Letters to the Editor

Improved "Impregnated Cathode"

ROBERTO LEVI
Philips Laboratories, Irvington-on-Hudson, New York
(Received February 10, 1955)

RECENTLY a new dispenser-type thermionic cathode, termed the Philips "impregnated cathode," has been announced[1] and some of its advantages over the "*L* cathode" described. Further development work has resulted in a greatly improved version of the original "impregnated cathode." The improved emitter has the following advantages over its predecessor.

First, the emission current density (measured pulse-wise by discharging periodically across the diode an *RC* network having a time constant of 60 microseconds at a repetition rate of 60 per second) has been increased by a factor of about four (see Fig. 1).

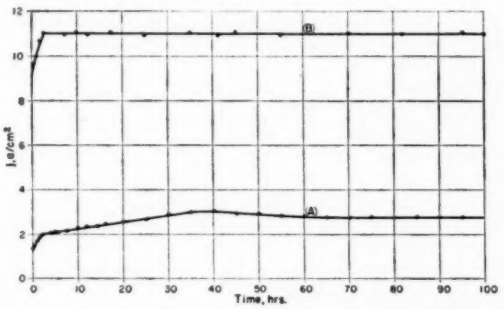


FIG. 1. Emission density *versus* aging time for (A) old impregnated cathode, (B) improved cathode. Aging temperature 1465°K; emission readings taken at 1410°K; cathode-anode separation =0.38 mm; applied voltage 1000 v.

Second, the long high-temperature aging process of about three days, previously required to obtain full and stable emission, has now been reduced to two or three hours (Fig. 1); if the tube is processed severely enough on the pumping system, the aging process may be entirely eliminated.

Third, the rate of total barium (Ba+BaO) evaporation from the new cathode, averaged over the first 2000 hours of life, is about 2×10^{-10} gram of Ba/cm² per second at 1465°K which is lower than the corresponding rate for the old type emitter at this operating temperature by a factor of four.[2] Since, however, the same emission current density can be obtained at lower operating temperature, the barium evaporation rate for equal emission density is further reduced relative to the old cathode by a factor of about eight.

Fourth, accelerated life tests at 1465°K have thus far exceeded 5000 hours without any failures (cathode thickness=1 mm). This time is of the order of magnitude of the life of the old cathode at the same temperature. However, if a comparison were to be made on the more significant basis of equal emission density, it is reasonable to expect at least an eightfold improvement in life for the new cathode.

The operating characteristics of the new emitter, as determined from a typical diode, are shown in Fig. 2. The zero-field emission determined from Schottky plots of these data yield the following values for the true (temperature dependent) work function and its temperature coefficient: $\phi=2.12$ ev at 1410°K, $\Delta\phi/\Delta T=3.2\times10^{-4}$ ev/degree. In computing the emission density from the observed current, it has been assumed that emission arises solely from the tungsten body (diameter=3.0 mm) with no contribution from the molybdenum support member (over-all diameter 3.3 mm).

The marked improvements enumerated above have been brought about by impregnating the porous tungsten body (porosity=17 percent) in a hydrogen atmosphere with a melt containing

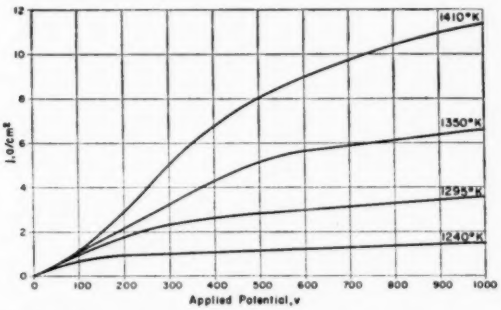


FIG. 2. *j-V-T* characteristics of planar diode. Cathode-anode separation =0.47 mm.

5 moles of BaO, 2 moles of Al_2O_3 and 3 moles of CaO. In the case of the old-type emitter, the impregnation was carried out in vacuum and the impregnant consisted of 5 moles of BaO and 2 moles of Al_2O_3.

The physical structure of the improved emitter is identical to that of the first type "impregnated cathode." Therefore, all of the latter's advantages[3] over the "*L* cathode" are retained.

[1] R. Levi, J. Appl. Phys. **24**, 233 (1953).
[2] We are indebted to W. C. Rutledge for this information.
[3] R. Levi, Le Vide **9**, No. 54, 284 (1954).

Electron Beam Method of Determining Density Profiles across Shock Waves in Gases at Low Densities*

DOUGLAS VENABLE AND DANIEL E. KAPLAN†
University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico
(Received January 3, 1955)

AS early as 1894, P. Lenard[1] demonstrated that the attenuation of cathode rays by gases depended strongly on the gas density. However, it has remained for later workers to apply this principle to the measurement of nonuniform density distributions. E. Schopper and N. Schumacher and A. E. Grün[2-4] have observed nontransient density distributions in gas jets by means of electron-beam shadowgraphs.

In the present work density profiles across shock waves propagating in gases were observed by means of changes in the transmission of an electron beam. The method has facilitated extending transient observations of shock phenomena to regions of low density where conventional methods (interferometric and x-ray techniques) do not respond well.

For the case where scattering follows an exponential law the transmission sensitivity, i.e., the rate of change of transmission with density, is a maximum when the electron path length through the gas is equal to the electron mean free path. This criterion defines the optimum particle energy. However, for this work the path length was permitted to be as much as several mean free paths. Transmission measurements in helium, neon, argon, and krypton demonstrated that path lengths as high as 5×10^{-5} g/cm² could be measured using 10-kev electrons.

The physical system consisted of a source of electrons which injected a colimated beam into the shock tube through a window. Diametrically opposite this window was another window through which electrons passed to the detector. An aperture lens system[5] collimated the electron beam. An *L*-cathode (Philips, type BP-1A) was chosen as an electron source because of its high emission and its resistance to poisoning. The window consisted of a 10- to 15-micron-diameter hole in one-mil platinum foil. Aluminized zapon foils, even when supported on 325-mesh gauze, proved unsatisfactory because they would withstand only a few shocks. In

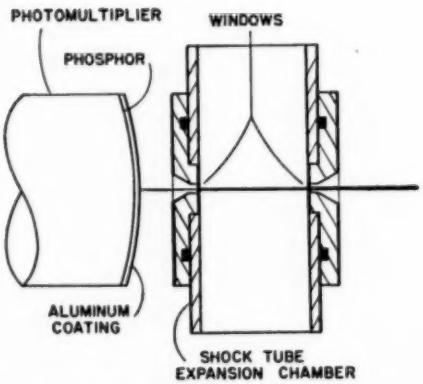


FIG. 1. Plan view of the experimental arrangement showing a section of the shock tube, the entrance and exit windows, and the detector. The heavy lines over the windows represent the platinum foil.

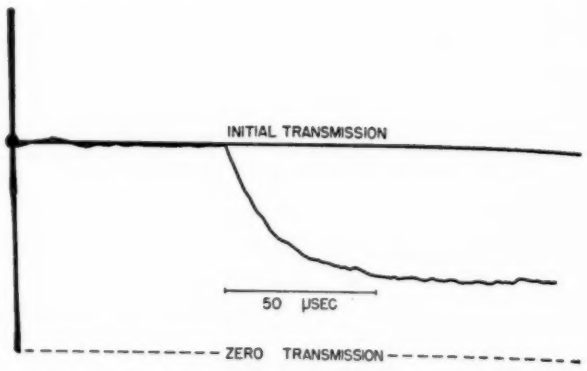


FIG. 3. A typical oscilloscope trace. The initial pressure of the argon in the expansion chamber was about 0.5 mm Hg. Helium at 580 mm Hg was used as a driving gas.

order to facilitate alignment of the beam and the two windows, the electron source was mounted on a ball-socket flange whose radius had its origin at the beam entrance window. The detector consisted of a No. 5819 photomultiplier whose face was coated with a terphenal activated polystyrene phosphor. An aluminum coating about 10 millimicrons thick was deposited on the phosphor and grounded to remove electrons. The detector output was presented on an oscilloscope. The combined time resolution of the detector and oscilloscope was less than one microsecond. Estimates indicated that the space resolution was less than 0.5 millimeter. Figure 1 shows the significant features of the system.

A shock was produced by rupturing a diaphragm which separated a high-pressure chamber from a low-pressure chamber. Each chamber was sufficiently long that on the basis of elementary theory[6] the duration of uniform flow behind the shock at the observation station was at least a hundred microseconds. The inside diameter of both chambers was $1\frac{1}{8}$ inches.

Although the injected beam current was constant during any observation, the currents varied by perhaps a factor of three among observations. Within this range the transmission was independent of beam current for densities above 5×10^{-7} g/cm^3. A transmission-density calibration curve is shown in Fig. 2 for 10-kev electrons in argon. These data were taken under static conditions of density. The error associated with the gas flow from the windows has not been evaluated.

Figure 3 shows a typical oscilloscope trace. The transmission-density calibration indicated that the final density was about that predicted[6] from the initial pressure ratio across the diaphragm. Here the rise time of the shock front to 90 percent of peak density is about 35 microseconds. It was observed that as the initial pressure in the expansion chamber was lowered to 0.05 mm Hg the rise time of the front increased to about 350 microseconds.

In this system a density profile represents an average density of the gas between the two windows. Thus a nonplanar shock or an undeveloped shock front could cause the density profile to have a long rise time such as observed in Fig. 3.

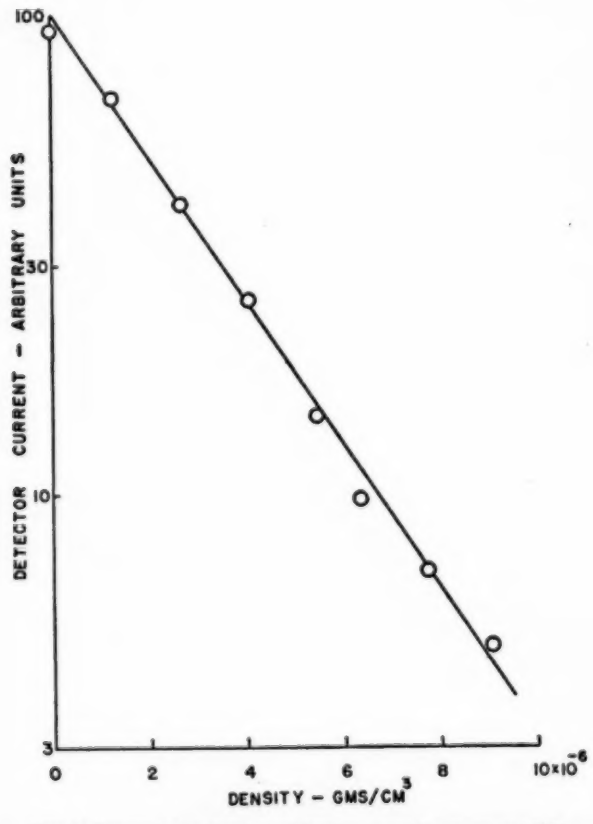


FIG. 2. Transmission-density calibration curve for argon at 25°C.

* Work done under the auspices of the U. S. Atomic Energy Commission.
† At present a graduate student at the University of California, Berkeley, California.

[1] P. Lenard, Ann. Physik **51**, 225–267 (1894).
[2] E. Schopper and B. Schumacher, Z. Naturforsch. **6a**, 700–705 (1951).
[3] B. Schumacher, Ann. Physik **13**, 404–420 (1953).
[4] Grün, Schopper, and Schumacher, J. Appl. Phys. **24**, 1527–1528 (1953).
[5] C. J. Davisson and C. J. Calbick, Phys. Rev. **42**, 580 (1932).
[6] F. W. Geiger and C. W. Mautz, "The shock tube as an instrument for the investigation of transonic and supersonic flow patterns," Report on Contract N6-ONR-232 of the Office of Naval Research, TO IV, Engineering Research Institute, University of Michigan, 1949.

Internal Friction of an Alloy of 16 Percent Aluminum in Iron*

C. WERT
University of Illinois, Urbana, Illinois
(Received December 20, 1954)

THE commercial use of the ferromagnetic alloys of iron and aluminum has lagged far behind the use of alloys of iron and silicon. This is so in spite of the fact that the higher electrical resistivity of the iron-aluminum alloys makes them more desirable for some purposes. The major reason for their lack of use is the difficulty in fabrication of rolled sheet. Recently there has been reported an extensive piece of work on the successful rolling of a sheet of an alloy of 16 wt. percent aluminum in iron (this alloy is called "16-Alfenol"[1]).

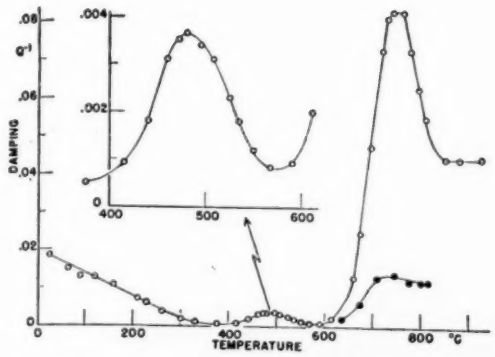


FIG. 1. Damping curves obtained for an alloy of "16 Alfenol." The open circles are for a specimen annealed at 1000°C prior to measurement. The inset shows an enlarged plot of the damping behavior around 500°C. The filled circles show the data obtained for a specimen annealed at 1350°C.

The present work reports some measurements of internal friction of an alloy of this composition. This investigation was first undertaken to see if such measurements could show the presence of interstitial impurities in quantities which could significantly affect the mechanical properties. No damping peaks were observed which could be identified as being associated with interstitial impurities. One must conclude then either that no interstitials were present or that the measurement was insensitive to them. There were observed, however, some interesting damping effects of different origins. They are described in this paper along with possible explanations of their origins.

One of the major problems was that of preparing the specimens for the torsion pendulum used to make the measurements. Because of the difficulty of working alloys of aluminum and iron in this compositoin range, the drawing of wires commonly used for the torsion pendulum presented a hard problem. This problem was avoided by the use of square wires which were machined from a sheet of the alloy that had been hot rolled to the proper thickness.† Four specimens were machined into wires about 1 mm×1 mm ×12 cm. A high temperature vacuum anneal served to relieve the cold work of rolling and machining and to establish a fairly uniform grain size. The measurements themselves were made in the vacuum apparatus described in a previous paper[2] (see Fig. 3, reference 2).

A typical measurement is shown in Fig. 1. The full curve (open circles) was obtained for a specimen that had been annealed at 1000°C for one hour. The data were taken with decreasing temperature after an additional anneal at 925°C in the apparatus following mounting of the sample. Three distinct regions of high damping were observed: one around 750°C, another around 475°C, and a a third below 300°C. The first two seem to be the result of relaxation processes with well-defined relaxation times. The last is not describable by a unique relaxation time.

The high damping which occurs around 750°C is almost surely the result of relaxation of stress in the grain boundaries. The basis for this conclusion lies in the following observations: The grain size of this sample was about 0.2 mm; the height of the peak was $Q^{-1}\cong 0.08$ at 750°C. A second sample which was annealed in vacuum at a much higher temperature than the first, one hour at 1350°C, showed a much lower damping a 750°C. In fact, the peak at 750°C was down to $Q^{-1}\cong 0.01$ for this sample, as is seen in Fig. 1 (curve passing through filled circles). Almost all the grains in this specimen extended completely across the 1 mm wire and almost all were from two to three mm long. Damping behavior of this type is characteristic of relaxation at grain boundaries, so it is likely that this is the grain boundary peak.

The damping peak at 475°C is not so easily identified. It is a rather low peak, Q^{-1} max$\cong 0.0035$, but has been reproduced at this level many times in these two specimens and also in a third specimen cut from a sheet of 14 percent aluminum in iron. In spite of number anneals and quenches neither its height nor position could be changed significantly.

The writer believes this peak to be caused by stress-induced ordering. The atomic-size difference between aluminum and iron is large enough to give rise to such a peak,[3] in fact the difference is so great that one might expect a peak much higher than $Q^{-1}\cong 0.0035$. However, this low value has a reasonable explanation: an alloy of 16 percent aluminum in iron exhibits long range order below 550°C. The damping peak lies well below this critical temperature, so the alloy must have a high degree of order at 475°C. Since a perfectly ordered alloy would show no damping from this source, it is reasonable to expect an alloy only slightly disordered to show only a small peak.

The damping that exists below 300°C is probably magnetic in origin and is the magnetoelastic static hysteresis.[4] It disappears completely above the Curie point, which is between 300°C and 400°C for the ordered alloy. It has no single relaxation time. Finally, the damping is strongly amplitude dependent, increasing linearly with strain for small strains. (The data in Fig. 1 are taken at the same strain amplitude for all temperatures.) This damping is not the same for all specimens and seemed highly sensitive to internal strains. To show it satisfactorily, the specimen had to be annealed in the apparatus after it was mounted.

* This work was supported in part by the Office of Ordnance Research, U. S. Army.

† The author is indebted to Mr. Pavlovic of the Westinghouse Corporation for this sheet of material.

[1] J. Nachmann and W. Buehler, J. Appl. Phys. **25**, 307 (1954).
[2] C. Wert, *The metallurgical use of anelasticity, research techniques in physical metallurgy* (Am. Soc. Metals, Cleveland, Ohio, 1953).
[3] C. Zener, Phys. Rev. **71**, 34 (1947).
[4] A. Nowick, Progr. Metal Phys. **4**, 64.

Optical Determination of Base Width in Grown *n-p-n* Silicon Crystals

ALAN FOWLER AND PASCAL LEVESQUE
Research Division, Raytheon Manufacturing Company, Waltham, Massachusetts

(Received December 13, 1954)

THE use of grown *n-p-n* silicon crystals to fabricate semiconductor devices has emphasized the need for an accurate measurement of the base width, i.e., the distance between the emitter and the collector junctions. Various methods of measuring base width have been reported, including scanning the base region with a fine probe[1] or light spot,[2] using alpha cut-off frequencies,[3] pulse delays,[4] and space charge widening.[5]

The technique presented here consisted of a sequence of etches on grown *n-p-n* silicon crystals to bring out the emitter and collector areas. This is an extension of work by Billig and Dowd[6] on grown *p-n* germanium junctions.

A bar of a grown *n-p-n* silicon crystal was potted in a thermo-

FIG. 1. Emitter junction of a grown *n-p-n* silicon crystal.

setting plastic with the leads from the emitter and collector free. After polishing, the sample was etched in freshly prepared CP−4[7] until the emitter junction was well delineated (Fig. 1). The sample was then etched electrolytically at room temperature in a 0.4 percent aqueous solution of potassium hydroxide using the collector as anode and the emitter as cathode. A current density of 10 amperes per cm² applied for ten to forty-five minutes revealed the collector junction (Fig. 2). The base width was verified by the deposition of barium titanate[8] on the two junctions (Fig. 3).

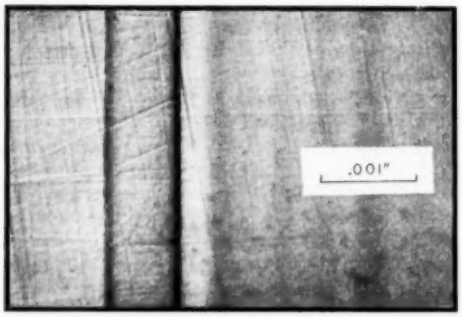


FIG. 2. Base width of a grown *n-p-n* silicon crystal.

FIG. 3. Barium titanate on junctions of a grown *n-p-n* silicon crystal.

In the units studied, the resistivity of the emitter was approximately one hundred times lower than that of the collector. A great difference in etching rate therefore should exist in the base-emitter interface. In fact, only a light etch with CP−4 was necessary to delineate clearly the beginning and end of the base-emitter transition region (Fig. 4).

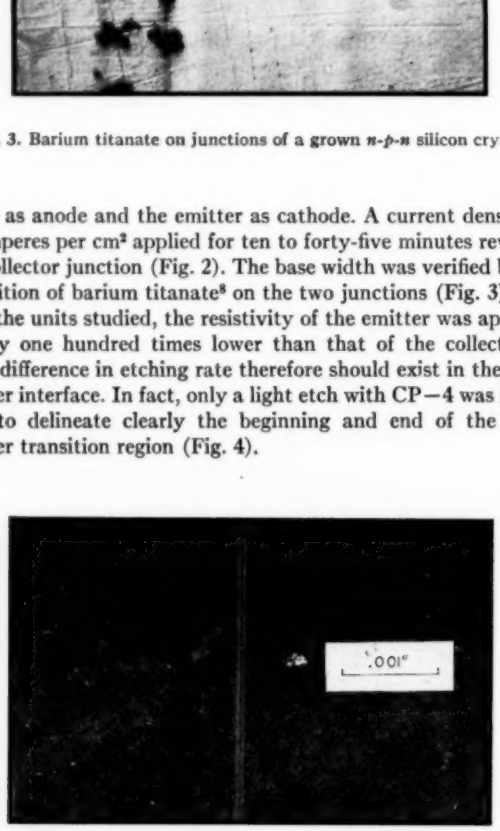


FIG. 4. Base-emitter transition region in grown *n-p-n* silicon crystal (sensitive tint).

A banded structure appeared in the emitter region to the right in Fig. 1. Similar structures in germanium have been reported,[9] and were attributed to sharp fluctuations in impurity concentration. This condition gave rise to a differential etch rate, which produced the banded structure observed. In general, the structure varied in intensity from one silicon crystal to another, and usually disappeared completely a few mils beyond the emitter-base interface.

The crystals used were grown by A. Adams, L. Bornstein, and W. Straub. The authors wish to acknowledge the valuable assistance of S. Cohen, L. Gerlach, and R. Mroczek during the course of this work.

[1] Pearson, Read, and Shockley, Phys. Rev. **85**, 1055 (1952).
[2] J. N. Shive, Proc. Inst. Radio Engrs. **40**, 1410 (1952).
[3] R. L. Pritchard, Proc. Inst. Radio Engrs. **40**, 1476 (1952).
[4] J. S. Schaffner, and J. T. Suran, J. Appl. Phys. **24**, 1355 (1953).
[5] H. Schenkel and H. Statz, Proc. Natl. Electronics Conference **10**, (1954).
[6] E. Billig and J. J. Dowd, Nature **172**, 115 (1953).
[7] J. R. Haynes and W. Shockley, Phys. Rev. **81**, 836 (1951).
[8] G. L. Pearson, U. S. Patent No. 2,669,692, February 16, 1954.
[9] P. R. Camp, J. Appl. Phys. **25**, 459 (1954).

Neutron Absorption in a Semi-Infinite Region

J. S. DWORK, *University of Vermont, Burlington, Vermont*
AND
H. HURWITZ, JR., *Knolls Atomic Power Laboratory,* Schenectady, New York*
(Received February 10, 1955)

THE Wiener-Hopf[1] method has been used to calculate the neutron absorption in a region 1 occupying the half-space $z<0$ when there is a uniformly distributed isotropic and monoenergetic neutron source in a region 2 which occupies the half-space $z>0$. The ratio p_1 of scattering cross section to total cross section in region 1 is assumed to differ from the corresponding ratio p_2 in region 2. The scattering in the two regions is assumed to be isotropic and not to change the neutron energy. The net flow of neutrons per unit time from region 2 to region 1 per unit area of interface is found to equal ΛS where S is the number of neutrons emitted by the source per unit time and volume in region 2 and Λ is given by the relationship

$$\Lambda=\lambda_2(1-p_1)\frac{f(p_2)-f(p_1)}{p_2-p_1}.$$

Here, λ_2 is the neutron mean free path in region 2 and

$$f(p)=\frac{1}{\pi}\int_0^{\pi/2}\frac{\log\left(\frac{1-px\cot x}{1-p}\right)}{\sin^2 x}dx.$$

It can be shown that

$$f(p)=\frac{1}{4}p+\left(\frac{1}{12}+\frac{1}{6}\log 2\right)p^2+\left(\frac{1}{24}+\frac{1}{3}\log 2-\frac{\pi^2}{96}\right)p^3+0(p^4)$$

as $p\to 0$, and

$$f(p)=\frac{1}{\sqrt{3}(1-p)^{\frac{1}{2}}}-0.710446+0(1-p)^{\frac{1}{2}} \quad \text{as} \quad p\to 1.$$

Thus, if $p_1=0$ (no scattering back from region 1 to 2),

$$\Lambda=\frac{\lambda_2}{4} \quad \text{if} \quad p_2=0,$$

$$\Lambda\cong L_2-0.71\lambda_2 \quad \text{if} \quad p_2\cong 1,$$

where

$$L_2=\lambda_2/(3[1-p_2])^{\frac{1}{2}}.$$

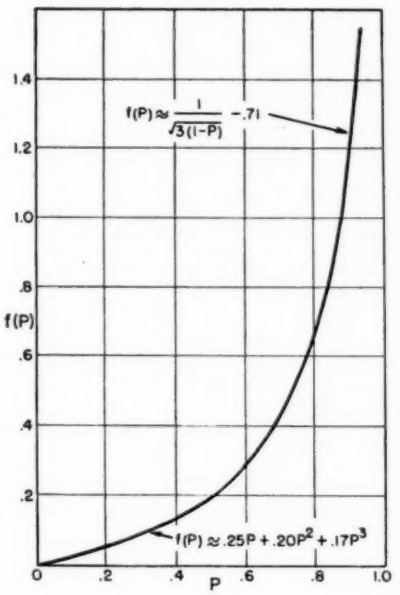


FIG. 1.

If $p_2=0$, the number of neutrons emerging from region 2 per unit time and area with direction vector in the solid angle, $d\Omega$ is proportional to $\cos\vartheta d\Omega$ where θ is the angle between the velocity vector and the z-axis. Thus, for $p_2=0$,

$$\Lambda=(\lambda_2/4)(1-A),$$

where A is the albedo of a semi-infinite slab for the case in which the incident neutron current has a cosine angular distribution. Thus[2]

$$(1-A)=4(1-p_1)f(p_1)/p_1.$$

If $p_1\cong 1$,

$$(1-A)\cong 4\left(\frac{1-p_1}{3}\right)^{\frac{1}{2}}.$$

This is the expression which would be obtained from diffusion theory.

A graph of $f(p)$ is given in Fig. 1.[3]

* Operated for the U. S. Atomic Energy Commission by the General Electric Company.

[1] For a description of a typical application of the Wiener-Hopf method to neutron transport problems, see G. Placzek and W. Siedel, Phys. Rev. **72**, 550 (1947). The presence of a source for $z>0$ introduces an additional term in the equation for the flux transforms which corresponds to Eq. (31) in this reference. It is possible to subtract a function from both sides of the equation which eliminates singularities arising from the source term so that the usual argument of analytical continuation can be applied. See also R. E. Marshak, Phys. Rev. **72**, 47 (1947).

[2] See Halpern, Lueneberg, and Clark, Phys. Rev. **53**, 173 (1938).

[3] This graph was prepared by D. M. Keaveney and S. R. Bigelow under the direction of J. H. Smith.

Erratum: On the Interaction of Two Similarly Facing Plane Shock Waves

[J. Appl. Phys. 25, 1549 (1954)]

C. A. FORD AND I. I. GLASS
Institute of Aerophysics, University of Toronto, Toronto, Canada

THE strength of the incident shock wave S_1 should read "$1.65a_1$" in the caption of Fig. 2.

Research Reports

This column will announce the availability of reports describing research in the fields of physics represented in the Journal of Applied Physics. The supplies of copies of most reports are strictly limited and are rarely reprinted when the initial printing is exhausted.

An announcement for this column should be fewer than 150 words in length and should provide information on the content, length, price, and address from which the report can be obtained. The announcement should be accompanied by a copy of the report, which will be reviewed to determine whether the report is of sufficient interest and relevance to warrant an announcement.

Theory and Application of Gyrator Networks

HERBERT J. CARLIN, Report R-355-53, PIB-289, March 18, 1954. Write to: Microwave Research Institute, Polytechnic Institute of Brooklyn, 55 Johnson Street, Brooklyn 1, New York. Price: $1.00.

This report presents a theoretical investigation of the circuit properties of linear passive but nonreciprocal networks. Networks containing gyrators (the basic nonreciprocal element) are first briefly analyzed. This is followed by a complete presentation of the synthesis of an assymmetric impedance matrix with complex elements. The class of networks having no impedance or admittance matrices (important in many microwave ferrite applications) is then treated by an extension of the synthesis technique in terms of scattering parameters, and the necessary and sufficient conditions for the realizability of any physical linear reciprocal or nonreciprocal system are given.

Structures having directional properties (biconjugate networks) are investigated as an application of the general synthesis techniques. The operation of ferrite devices employing directional elements such as circulators, duplexers, one-way lines, frequency separators, and wide-band filters is treated from a networks point of view. Equivalent circuit representations are given.

Basic Circuitry of the MIDAC and MIDSAC

Report No. 1947-2-T, 116 pages. Write to: University of Michigan Press, Ann Arbor, Michigan.

Digital computer design and circuitry for two large-scale machines, the general purpose MIDAC and the special purpose MIDSAC, are presented in this report by the Willow Run Research Center of the University of Michigan.

Books Reviewed

Development of the Guided Missile. KENNETH W. GATLAND. Pp. 292. Second edition, Philosophical Library, Inc., New York, 1954. Price $4.75.

This book presents a survey of the published facts on rockets and guided missiles for both military and peaceful purposes. It includes discussions of air-to-air missiles, guided bombs, long-range missiles, rockets for high-altitude research, satellites, and interplanetary travel. The introduction considers military and economic feasibility of guided missiles as compared to conventional weapons. Appendixes describe the all-important telemetering systems and give photographs to scale of 40 rockets from various countries. The appendix which appeared in the first edition, giving a table of characteristics of missiles, has been enlarged to include 140 rockets from eight countries. The author, who is an aircraft design engineer and founder-member of the British

Interplanetary Society, has included a bibliography briefly describing other books on rocket development and space travel.

Magnetic Amplifiers. H. F. STORM. Pp. 545+xix. John Wiley and Sons, Inc., New York, 1955. Price $13.50.

The purpose of this book is to explain the operation of static magnetic amplifiers and to illustrate their principal uses. It is intended for electrical engineers and senior students in this field. Although interest is centered on the amplifier as such, several chapters are devoted to a modern theory of magnetism, characteristics of magnetic materials, magnetic testing, manufacture of saturable reactors, and properties of metallic rectifiers. The subjects of peaking transformers, voltage stabilizers, nonlinear resonance circuits, and reactors for large mechanical rectifiers are treated in a separate chapter. In general, mathematical complexity is avoided and equations are often graphically illustrated. Applications of magnetic amplifiers are contained in six chapters. The bibliography covers 901 publications and patents for the period between 1887 and 1951.

Multipole Fields. M. E. ROSE. Pp. 99+viii. John Wiley and Sons, Inc., New York. Price $4.95.

The theme of this book is the angular-momentum and parity or transformation properties of the multipole fields. Angular-momentum properties form the basis of description of a multitude of physical systems. In this volume, the author has utilized group theoretic methods in describing the electromagnetic field. Although no extensive knowledge of group theory is assumed, a general familiarity with quantum mechanics and electromagnetic theory is needed. The necessary basis of Maxwell theory is briefly presented at the outset. Multipole field theory is developed in the first three chapters. The last three are devoted mainly to applications. The specific topics treated are classical field equations, multipole fields, properties of the multipole fields, retarded electromagnetic interaction, internal conversion, and emission of gamma radiation.

Books Received

Electricity and Magnetism. RALPH P. WINCH. Pp. 755+xii, with Figs. Prentice-Hall, Inc., New York, 1955. Price $7.75.

Tables of the Error Function and Its Derivative. National Bureau of Standards Applied Mathematics Series 41. Pp. 302. Price $3.25.

Tables of Functions and of Zeros of Functions. National Bureau of Standards Applied Mathematics Series 37. Pp. 211. Price $2.25.

Abstracts of the Literature on Semiconducting and Luminescent Materials and Their Applications. The Electrochemical Society. Pp. 169. John Wiley and Sons, Inc., New York, 1955. Price $5.00.

Multipole Fields. M. E. ROSE. Pp. 99+viii. John Wiley and Sons, Inc., New York, 1955. Price $4.95.

Advanced Mathematics for Engineers. H. W. REDDICK AND F. H. MILLER. Pp. 548+xiv, Figs. 131. John Wiley and Sons, Inc., New York, 1955. Price $6.50.

Elements of Probability Theory. HARALD CRAMER. Pp. 281, Figs. 29. John Wiley and Sons, Inc., New York, 1955. Price $7.00.

Titanium in Iron and Steel. GEORGE F. COMSTOCK. Pp. 294 +xii, with Figs. John Wiley and Sons, Inc., New York, 1955. Price $6.00.

Advances in Electronics and Electron Physics. Edited by L. MARTON. Pp. 538+xi, with Figs. Academic Press, Inc., New York, 1955. Price $11.80.

Announcements

Course in Modern Industrial Spectrography at Boston College

Boston College has announced a special two-week intensive course in Modern Industrial Spectrography at Chestnut Hill, Boston, Massachusetts, from July 11 to July 22. The course is particularly designed for chemists and physicists from industry who wish to learn the techniques of emmission spectroscopy as an analytical tool. An optional third week of laboratory practice will be made available to those who desire it. Information on the course may be obtained from Professor James J. Devlin, Physics Department, Boston College, Chestnut Hill 67, Masachusetts.

Summer Course in Electron Microscopy

The Cornell Summer Laboratory Course in Techniques and Applications of the Electron Microscope will be given again this year from June 13 to June 25. Professor Benjamin M. Siegel will again be the director, and guest lecturers this year will be Professor Cecil E. Hall (M.I.T.) and Professor Robley C. Williams (University of California, Berkeley). The course is designed to give an intensive survey of basic theory and interpretation of results. The registration is limited to a small group so that ample facilities are available for each individual to pursue laboratory work in his special field at an introductory or advanced level. Inquiries should be addressed to Professor B. M. Siegel, Department of Engineering Physics, Rockefeller Hall, Cornell University, Ithaca, New York.

Cover Photograph

The photograph shows a copper single crystal sphere which has been oxidized. This $\frac{5}{8}$-in. sphere was cut from a single crystal rod of 99.999 percent pure copper. The sphere was polished mechanically and electrolytically in order to remove the strained layer resulting from cutting and to obtain a very smooth surface. Then the sphere was annealed in hydrogen at 500°C and oxidized in oxygen at 250°C for 45 min. The different crystal faces of copper oxidize at different rates, thereby building up oxide layers of different thicknesses. These oxide films exhibit interference colors, and the approximate thickness of the oxide film can be determined from these colors. Thus the (311) face shows first order red, ~250 A, while the (100) face shows fifth order green, ~5000 A. It is a simple matter to pick out the various faces on the crystal. The center of the cross is a (100) face; the 4 arms of the cross extend toward (110) faces; the triangular area between two (110) faces is a (111) face. The single crystal sphere is useful in studying other types of surface reactions such as catalysis, electrochemical phenomena, friction and wear, wetting, and corrosion. In general the surface properties of metals vary greatly with crystal face.

The cover photograph and this explanation were furnished by Dr. F. W. Young, Jr., of the University of Virginia.